2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Deep Interactive Full Transformer Framework for Point Cloud Registration

Guangyan Chen[1], Meiling Wang[1], Qingxiang Zhang[1], Li Yuan[2], Tong Liu[1], Yufeng Yue[1*]

Abstract— **Point cloud registration is a crucial technology in the fields of robotics and computer vision. Despite the significant advances in point cloud registration enabled by Transformer-based methods, limitations persist due to indistinct feature extraction, noise sensitivity, and outlier handling. These limitations stem from three factors: (1) the inefficiency of convolutional neural networks (CNNs) to capture global relationships due to their local receptive fields, resulting in extracted features susceptible to noise; (2) the shallow-wide architecture of Transformers, coupled with a lack of positional information, leading to inefficient information interaction and indistinct feature extraction; and (3) the omission of geometrical compatibility leads to ambiguous identification of incorrect correspondences. To overcome these limitations, we propose the Deep Interactive Full Transformer (DIFT) network for point cloud registration, which consists of three key components: (1) a Point Cloud Structure Extractor (PSE) for modeling global relationships and retrieving structural information; (2) a Point Feature Transformer (PFT) for establishing comprehensive associations and directly learning the relative positions between points; and (3) a Geometric Matching-based Correspondence Confidence Evaluation (GMCCE) method for measuring spatial consistency and estimating correspondence confidence. Experimental results on ModelNet40 and 3DMatch datasets demonstrate the superior performance of our proposed method compared to existing state-of-the-art methods. The code for our method is publicly available at `https://github.com/CGuangyan-BIT/DIFT`.**

I. Introduction

Point cloud registration is a crucial technique in robotics and computer vision that estimates a rigid transformation for aligning two point clouds. In recent decades, learning-based point cloud registration methods have developed from the convolutional neural network (CNN)-based methods [1]–[3], to recent Transformer-based methods [4]–[7]. CNN-based methods integrate CNNs to extract point-wise features and establish point-to-point correspondences based on feature similarity. However, since they extract features from each point cloud separately, they have difficulty identifying common structures between two point clouds and extracting discriminative features.

Attention-based models, such as Transformers [8], have demonstrated outstanding performance in natural language processing (NLP) tasks [9]–[11], and have been shown to excel in feature extraction and information aggregation in computer vision tasks [12]–[18]. Inspired by the success of Transformers, recent works [4]–[6], [19], [20] investigated the integration of Transformer models. While point cloud registration shares many similarities with other point cloud processing tasks (e.g.: object detection, semantic segmentation), two crucial distinctions set it apart. First, while other tasks operate on a unified coordinate system for processing point clouds, registration must deal with different coordinate systems. Second, other tasks typically cluster points into predefined categories, while registration focuses on minimizing the differences between point pairs and distinguishing each pair from all other points. These differences require registration methods to extract more representative and discriminative features.

The majority of recent registration methods [4]–[6], [19] employ the attention mechanism to establish inter-point-cloud associations for information aggregation, thereby allowing one point cloud to be aware of the other. However, substantial gaps remain in modeling global relations, enhancing feature richness, and detecting inliers: (1) current methods mainly utilize CNNs for single point cloud feature extraction, which leads to the sensitivity to noise due to the local receptive fields of CNNs; (2) the insufficient associations established by shallow-wide Transformers and lack of positional information prevent the model from enhancing feature richness and extracting distinct features; (3) the inlier detection modules neglect spatial consistency, resulting in a large proportion of outlier matches being preserved.

Motivated by the limitations of previous Transformer-based methods, we propose a novel full Transformer framework named Deep Interactive Full Transformer (DIFT), which capitalizes on the Transformer architecture to achieve a global receptive field and deep information interaction. The process of deep information interaction improves the discrimination of the extracted features significantly. In summary, the main contributions are four-fold: (1) A Point Cloud Structure Extractor (PSE) is proposed to model global relations and integrate structural information. (2) A Point Feature Transformer (PFT) is proposed to increase the richness of feature representation. (3) A Geometric Matching-based Correspondence Confidence Evaluation (GMCCE) method is proposed to estimate the correspondence confidence based on geometric constraints. (4) DIFT is compared with extensive registration methods on ModelNet40 and 3DMatch, achieving superior performance.

The remainder of this paper proceeds as follows: Sec. II reviews the related work. Sec. III introduces the framework and each module of DIFT. Sec. IV presents the experimental procedures and results. Sec. V offers concluding remarks.

This work is partly supported by National Natural Science Foundation of China under Grant 62003039, 62233002, and the CAST program under Grant No. YESS20200126. (Corresponding Author: Yufeng Yue, yueyufeng@bit.edu.cn)

[1]Guangyan Chen, Meiling Wang, Qingxiang Zhang, Tong Liu, and Yufeng Yue are with the School of Automation, Beijing Institute of Technology, Beijing, 100081, China.

[2]Li Yuan is with the School of Electrical and Computer Engineering at Peking University & Pecheng Lab, Shenzhen, 518055, China.

II. Related Work

A. CNN-Based Registration Methods

The success of deep learning in point cloud processing tasks [21]–[24] enables its application in point cloud registration. One pioneering work is PointNetLK [3], which extracts global features using PointNet [25] and applies the inverse compositional Lucas-Kanade (IC-LK) algorithm [26] to align two point clouds. PointNetLK Revisited [2] was proposed to improve the numerical instabilities of PointNetLK using analytical Jacobians. However, since PointNet cannot aggregate the information from two point clouds, they are sensitive to partially visible point clouds. Deep Gaussian mixture registration (DeepGMR) [27] relies on a neural network to predict the GMM parameters and recover the optimal transformation. However, due to the independence of the feature extraction from two point clouds, the features extracted by DeepGMR are indistinct. The robust point matching network (RPM-Net) [28] is proposed to apply the Sinkhorn [29] method to establish soft correspondences from hybrid features, thereby enhancing the robustness to noise. In summary, these methods extract features from each point cloud separately and lack information interaction between two point clouds, which are inefficient in discriminative feature extraction and contextual information aggregation, especially in partial-to-partial point cloud registration tasks.

B. Transformer-Based Registration Methods

Inspired by the success of Transformers in NLP and computer vision, researchers have begun to utilize Transformers to extract contextual information between two point clouds. Deep closest point (DCP) [4] extracts features using dynamic graph CNN (DGCNN) [30] and utilizes the Transformer [8] to aggregate information. However, DCP lacks an overall understanding of the point cloud because of its local receptive field, which leads to sensitivity to noise. A multiplex dynamic graph attention network (MDGAT) [5] dynamically constructs a multiplex graph based on an attention mechanism. A geometry guided network [7] encodes features with a fully connected graph based on the self-attention mechanism. Robust graph matching (RGM) method [6] adopts a Transformer to aggregate information along soft graph edges. Recent registration Transformer (REGTR) [31], utilizes attention layers to generate correspondences directly. In summary, these methods mainly focus on modeling local relations by convolution encoders, which restricts their ability to model global relations. Furthermore, the information interaction is inefficient due to the shallow-wide Transformer architecture and the lack of positional encoding.

III. The Proposed Deep Interactive Full Transformer

Given two point clouds $X = \{x_1, x_2, ..., x_N\} \subseteq R^3$ and $Y = \{y_1, y_2, ..., y_M\} \subseteq R^3$, which are denoted by src and tgt, respectively. The objective of DIFT is to recover a transformation consisting of a rotation matrix $R \in SO(3)$ and a translation vector $t \in R^3$ that aligns X to Y.

The overall DIFT pipeline is shown in Fig. 1. In the training phase, the pipeline begins with extracting pointwise features $\mathcal{F}_X$ and $\mathcal{F}_Y$ from src and tgt separately using PSE. Then, deep information interaction is conducted by PFT to learn contextual information and extract discriminative features Φ_X and Φ_Y. These features are matched to establish putative correspondences $M\{x_i, y_j\}$. Finally, the weighted Procrustes module estimates the optimal transformation $\{R, t\}$ based on the established correspondences M and the similarity $\mathcal{S}$ between the corresponding feature vectors $\{\Phi_{x_i}, \Phi_{y_j}\}$. During testing, the GMCCE module is introduced to evaluate the correspondence confidence $\widetilde{C}$, then the weighted Procrustes module estimates the optimal transformation based on the confidence $\widetilde{C}$.

A. Point Cloud Structure Extractor

Since the previous Transformer-based methods mainly employ features from CNN, which cannot model global relations. Therefore, the PSE module is designed to enhance the robustness to noise by modeling dependencies in the entire point cloud. Fig. 1(a) shows the PSE architecture, which consists of two components: Local Feature Integrator (LFI) and Transformer encoder [8].

The limitations of Transformers in extracting structural features [16], [17] and the lack of semantic information in point coordinates pose significant challenges, including the difficulty of convergence and the requirement for large quantities of training data. To overcome these challenges, the LFIs are designed to progressively structurize the point cloud. The n_{th} LFI layer ($n = 1, \ldots, N_l$) searches for the graph $\mathcal{G}_n$ that contains the features F_n of $\mathcal{K}$ nearest points for each point in the point cloud, where N_l denotes the number of LFI layers. Then, the n_{th} LFI layer integrates structural information by concatenating (Concat) nearby points in the graph $\mathcal{G}_n$ to construct integrated feature vectors f_n. Specifically, the LFI method applies the k-nearest neighbors (KNN) algorithm in geometric space, as opposed to feature space, providing two benefits: (1) introducing convolutional inductive bias for fast convergence; (2) reducing computational expense significantly.

The extracted features f_n are then passed through a Transformer encoder to model global relations [32]. Each Transformer encoder layer consists of a multilayer perceptron (MLP), layer normalization (LN), and a multi-head self-attention operation (MSA). The encoder obtains F_{n+1} as

$$\begin{aligned} \hat{f} &= \mathrm{LN}(\mathrm{MSA}(f_n)) + f_n), \\ F_{n+1} &= \mathrm{LN}(\mathrm{MLP}(\hat{f})) + \hat{f}), \end{aligned} \tag{1}$$

where F_{n+1} is the input of the $n+1_{\text{th}}$ LFI layer; MSA is based on multi-head attention (MA), MSA and MA are defined below:

$$\begin{aligned} \mathrm{MSA}(f_n) &= \mathrm{MA}(f_n, f_n, f_n), \\ \mathrm{MA}(F_Q, F_K, F_V) &= \mathrm{Concat}(A_1, \ldots, A_h)W^O, \\ \mathrm{Att}(Q, K, V) &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d_K}})V, \end{aligned} \tag{2}$$

where A_i represents the output of the attention function Att,

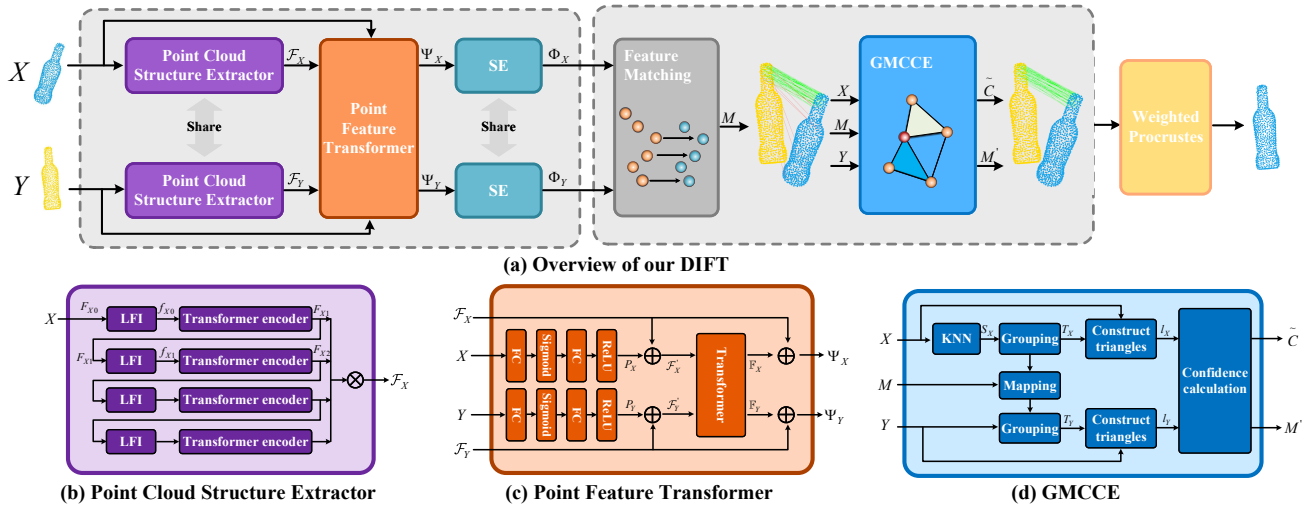

Fig. 1. (a) The network architecture of the Deep Interactive Full Transformer (DIFT). DIFT consists of three main components: (b) a Point Cloud Structure Extractor (PSE) and (c) a Point Feature Transformer (PFT) to extract features, where $\otimes$ denotes concatenation, $\oplus$ denotes matrix addition; (d) Geometric Matching-based Correspondence Confidence Evaluation (GMCCE) module to evaluate correspondence confidence.

which takes the linearly projected queries Q, keys K, and values V as input ($i = 1, ..., h$). The dimensionality of keys K is denoted as d_K. By performing h attention functions in parallel, MSA effectively captures the relationships between all points. Specifically, each attention function Att generates an attention map through scaled dot-product, which serves to aggregate information from the entire point cloud by multiplying the map by V. The results of each Att are then concatenated and projected to obtain the final values.

Finally, obtaining all output values F_{n+1} of Transformer encoders, low-order and high-order features are merged by concatenating the output values of each layer as

$$\mathcal{F} = \mathrm{LN}(\mathrm{ReLU}(\mathrm{Concat}(F_2, F_3, \cdots, F_{N_l+1}))). \tag{3}$$

B. Point Feature Transformer

The discriminative power of the extracted features $\mathcal{F}_X$ and $\mathcal{F}_Y$ is limited as they are independent of each other. To address this issue, a Point Feature Transformer (PFT) is designed to learn the contextual information of two point clouds and extract distinct features. Fig. 1(b) shows the architecture of the PFT, which consists of the Transformer based encoder-decoder and the positional encoding network.

To learn the relative position between points directly, a positional encoding network consisting of fully connected layers FC, rectified linear unit ReLU activation, and sigmoid activation is introduced, the positional encoding network extracts positional information P_X, P_Y as

$$\begin{aligned} P_X &= \mathrm{ReLU}(\mathrm{FC}(\mathrm{Sigmoid}(\mathrm{FC}(X)))), \\ P_Y &= \mathrm{ReLU}(\mathrm{FC}(\mathrm{Sigmoid}(\mathrm{FC}(Y)))). \end{aligned} \tag{4}$$

Subsequently, positional information P_X, P_Y are incorporated to the features $\mathcal{F}_X$, $\mathcal{F}_Y$, yielding the updated features $\mathcal{F}_X^{'}$, $\mathcal{F}_Y^{'}$, respectively. To aggregate information from src and tgt simultaneously, a standard Transformer ϕ is adopted, which consists of an encoder (Eq. 1) and a decoder. The Transformer decoder consists of a multi-head cross-attention operation (MCA), in addition to MSA (Eq. 2), MLP, and LN. Taking $\phi(\mathcal{F}_Y^{'}, \mathcal{F}_X^{'})$ as an example, the procedure of the decoder is defined as

$$\begin{aligned} \mathcal{F}_X^S &= \mathrm{LN}(\mathrm{MSA}(\mathcal{F}_X^{'}) + \mathcal{F}_X^{'}), \\ \mathcal{F}_X^C &= \mathrm{LN}(\mathrm{MCA}(\mathcal{F}_Y^S, \mathcal{F}_X^S) + \mathcal{F}_X^S), \\ \mathbb{F}_X &= \mathrm{LN}(\mathrm{MLP}(\mathcal{F}_X^C) + \mathcal{F}_X^C), \end{aligned} \tag{5}$$

where $\mathrm{MCA}(\mathcal{F}_Y^S, \mathcal{F}_X^S) = \mathrm{MA}(\mathcal{F}_X^S, \mathcal{F}_Y^S, \mathcal{F}_Y^S)$ (Eq. 2); features F_X^S are obtained based on MSA, and features $\mathcal{F}_Y^S$ are acquired through the encoder, then the attention map is generated using MCA, facilitating the establishment of associations between points across src and tgt. This enables $\mathcal{F}_X^S$ to receive information from $\mathcal{F}_Y^S$ and improve the discrimination of $\mathcal{F}_X^S$.

The preceding methods' employment of a shallow-wide architecture resulted in limited associations for information interaction, consequently causing low feature richness. In contrast, this paper employs a deep-narrow architecture to establish comprehensive associations. Overall, the features Ψ_X and Ψ_Y generated by Transformers are formulated as

$$\begin{aligned} \Psi_X &= \mathcal{F}_X + \phi(\mathcal{F}_Y^{'}, \mathcal{F}_X^{'}), \\ \Psi_Y &= \mathcal{F}_Y + \phi(\mathcal{F}_X^{'}, \mathcal{F}_Y^{'}). \end{aligned} \tag{6}$$

To adaptively recalibrate the channel-wise features based on their contribution to registration, a squeeze-and-excitation (SE) module [33] is adopted. The SE module initially generates a channel descriptor by employing average pooling on the input features and subsequently converts it into channel weights using a neural network. Finally, the input features are rescaled with the channel weights to yield rescaled features Φ_X and Φ_Y.

Given features Φ_X and Φ_Y, a set of putative correspondences $M \in R^{N\times 2}$ are established by finding the most similar features Φ_{yj} for Φ_{xi}. To demonstrate the discrimination of the extracted features Φ_X and Φ_Y, an intuitive example is shown in Fig. 2, t-SNE visualization (a) clearly shows that the points in each pair are located in a similar area and distinguished from other points, enabling DIFT to represent

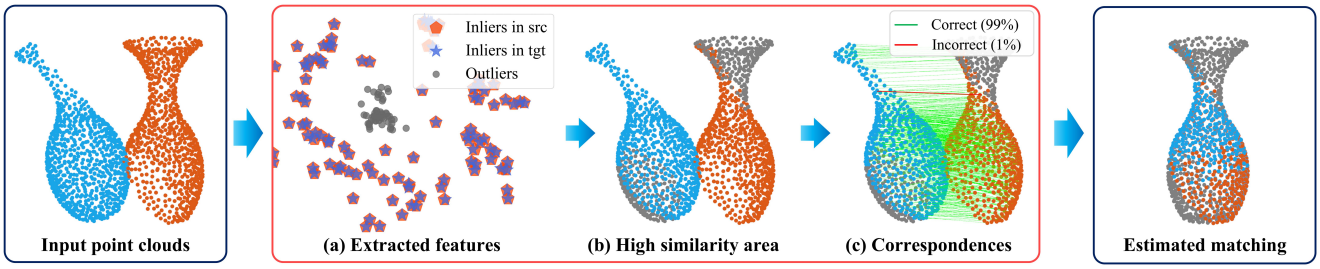

Fig. 2. An illustrative case of the extracted features. (a) The t-SNE visualization [34] of extracted features shows that DIFT obtains the discriminative features, which enables DIFT to (b) identify the overlapping area and (c) establish reliable correspondences.

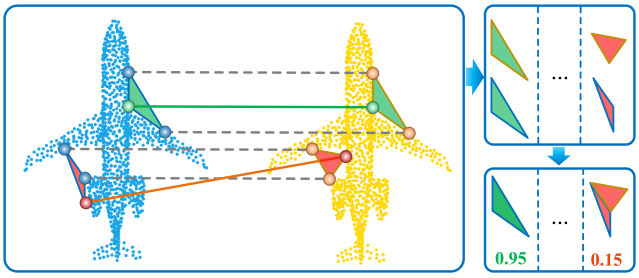

Fig. 3. The demonstration of the triangulated descriptor. When the correspondence $\{x_i, y_j\}$ is correct, the corresponding triangles are similar, and the confidence $\widetilde{C}(x_i, y_j)$ is high. Otherwise, $\widetilde{C}(x_i, y_j)$ decreases to a small value if $\{x_i, y_j\}$ is incorrect.

overlapping regions with (b) high similarity area and (c) establish accurate correspondences in overlapping regions, consequently, two point clouds are registered precisely.

C. Geometric Matching-Based Correspondence Confidence Evaluation

Since DIFT extracts discriminative features and establishes accurate correspondences in the high similarity area, leveraging the features Φ_X and Φ_Y enables the search for a correspondence subset with a higher inlier ratio. The subset is generated by selecting the $\mathbb{N}$ correspondences with the highest similarity to provide reliable correspondences. Additionally, filtering out any incorrect correspondences within the generated subset can further improve accuracy. Therefore, the GMCCE module is designed to evaluate correspondences with a triangulated descriptor. As shown in Fig. 3, the descriptor employs the side length of triangles to capture geometric characteristics, which offers two advantages: (1) expressing length and angle simultaneously; (2) establishing connections between sampled points.

The GMCCE module is presented in Fig. 1(c), to better illustrate the GMCCE, we detail the evaluation procedure for the putative correspondence $\{x_i, y_j\}$. First, KNN is performed to search for $\mathbb{K}$ sampled points S_X of x_i in *src*, then doublets $D_X \in R^{\mathbf{K}\times 2\times 3}$ are generated by combining sampled points S_X in pairs, where $\mathbf{K} = \binom{\mathbb{K}}{2}$. Afterward, triplets $T_X \in R^{\mathbf{K}\times 3\times 3}$ are obtained by combining D_X and x_i, specifically, each triplet contains x_i and a doublet in D_X. Subsequently, triplets T_Y are acquired by mapping T_X in accordance with the correspondences M. Then, GMCCE calculates the lengths l_X, $l_Y \in R^{\mathbf{K}\times 3}$ of the triplets T_X and T_Y, respectively. Afterward, the overall error $\mathcal{E}(x_i, y_j)$ is calculated by summing the $\mathbb{K}$ smallest length errors $\mathcal{L}$ as

$$\begin{aligned}\mathcal{E}(x_i, y_j) &= \sum \mathrm{Mink}([\mathcal{L}(T_X^1, T_Y^1), \cdots, \mathcal{L}(T_X^{\mathbf{K}}, T_Y^{\mathbf{K}})]),\\ \mathcal{L}(T_X^\beta, T_Y^\beta) &= \sqrt{\sum_{i=1}^{3}(l_X^{\beta i} - l_Y^{\beta i})^2 / \sum_{i=1}^{3}(l_X^{\beta i} + l_Y^{\beta i})^2},\end{aligned} \tag{7}$$

where $l^{\beta i}$ denotes i-th edge lengths of the triangles constructed by T^β; Mink is the operation of taking the $\mathbb{K}$ smallest values. Finally, the confidence $\widetilde{C}$ is evaluated as

$$\widetilde{C}(x_i, y_j) = \psi(2 \times \mathrm{sigmoid}(-\lambda \mathcal{E}(x_i, y_j))), \tag{8}$$

where λ is the parameter to adjust the sharpness of the confidence evaluation; ψ is the filter to filter out correspondences with confidence smaller than τ.

D. Loss Functions and Details

The loss functions to train our DIFT are defined below:

Transformation loss: L_t measures the error between the predicted motion R_{XY}, t_{XY} and ground-truth motion R^*_{XY}, t^*_{XY} from X to Y:

$$L_t = \|R_{XY}^T R^*_{XY} - I\|^2 + \|t_{XY} - t^*_{XY}\|^2. \tag{9}$$

Discrimination loss: L_d measures the discriminative power of extracted features and the accuracy of established correspondences:

$$\begin{aligned}L_d = -\frac{1}{\|M\|} \sum_{((x_i, y_j)\in M)} &[C(x_i, y_j) \times \ln \mathcal{S}(x_i, y_j)\\ &+ (1 - C(x_i, y_j)) \times \ln(1 - \mathcal{S}(x_i, y_j))],\end{aligned} \tag{10}$$

where $C(x_i, y_j) = 1$ if the correspondence $\{x_i, y_j\}$ is correct; otherwise, $C(x_i, y_j) = 0$. $\mathcal{S}(x_i, y_j)$ denotes the similarity between the corresponding features Φ_{xi} and Φ_{yj}.

Cycle consistency loss: L_c measures the consistency between the predicted motion from X to Y and Y to X:

$$L_c = \|R_{XY} R_{YX} - I\|^2 + \|t_{XY} + t_{YX}\|^2. \tag{11}$$

Implementation Details: Each LFI layer searches for the graph $\mathcal{G}_n$ with $\mathbb{K} = 20$ points and the Transformer encoder outputs features with 64 dimensions. In the GMCCE module, the parameters are set to $\mathbb{N} = 400$, $\mathbb{K} = 10$, $\lambda = 90$, and $\tau = 0.6$. In addition, DIFT is trained using Adam [39] with an initial learning rate 3e-5.

TABLE I

PERFORMANCE ON MODELNET40 DATASET. SCENE1, SCENE2, SCENE3 REPRESENT CLEAN POINT CLOUDS, LOW NOISE PARTIAL-TO-PARTIAL POINT CLOUDS, HIGH NOISE PARTIAL-TO-PARTIAL POINT CLOUDS RESPECTIVELY. THE THREE BEST RESULTS ARE HIGHLIGHTED IN **RED**, **GREEN**, **BLUE**.

Method	Reference	Scene 1 (clean)				Scene 2 (low-noise partial)				Scene 3 (high-noise partial)			
		R_{RMSE}	R_{MAE}	t_{RMSE}	t_{MAE}	R_{RMSE}	R_{MAE}	t_{RMSE}	t_{MAE}	R_{RMSE}	R_{MAE}	t_{RMSE}	t_{MAE}
ICP [35]	SPIE 1992	25.09	14.15	0.157	0.106	20.12	11.24	0.13	0.092	20.05	11.33	0.13	0.09
PNetLK [3]	CVPR 2019	12.02	4.954	0.0064	0.0038	18.10	12.38	0.131	0.101	18.33	12.17	0.14	0.108
DCP_V2 [4]	CVPR 2019	3.242	2.076	0.0024	0.0015	4.43	2.92	0.029	0.022	12.29	7.84	0.097	0.078
DeepGMR [27]	ECCV 2020	0.023	0.016	3e-5	2e-5	7.15	4.84	0.13	0.107	8.957	6.243	0.154	0.128
IDAM [36]	ECCV 2020	1.59	1.109	0.0259	0.018	14.44	8.54	0.10	0.07	18.91	11.78	0.093	0.067
PNetLK_R [2]	CVPR 2021	1.385	0.120	0.0085	0.0006	4.86	2.21	0.062	0.032	6.224	2.495	0.093	0.033
Reagent [37]	CVPR 2021	1.073	0.939	0.0023	0.0020	9.37	8.22	0.055	0.043	11.85	10.47	0.063	0.05
RGM [6]	CVPR 2021	**1.8e-4**	**1.2e-5**	**2.7e-6**	**1.6e-7**	**0.741**	**0.099**	**2.4e-3**	**8.1e-4**	**2.068**	**0.633**	**0.016**	**0.0061**
RIENet [38]	CVPR 2022	**0.012**	**0.009**	**4.1e-6**	**2.6e-6**	**0.554**	**0.096**	**7.7e-4**	**2.9e-4**	11.85	1.26	**0.018**	**0.0035**
RegTR [31]	CVPR 2022	1.037	0.323	0.0082	0.0026	1.605	0.665	0.016	0.0083	**1.909**	**0.771**	**0.018**	0.0065
Ours	-	**2.3e-6**	**1.5e-6**	**1.7e-8**	**1.1e-8**	**0.014**	**0.010**	**6.7e-5**	**5.3e-5**	**0.593**	**0.286**	**0.0056**	**0.0021**

IV. EXPERIMENTAL RESULTS

A. Matching and Registration Performance on ModelNet40

ModelNet40: The proposed algorithm and baseline methods are evaluated on ModelNet40 [40]. This dataset includes 12,311 meshed computer-aided design (CAD) models in 40 categories, of which 80% are designated for training and the remaining are designated for testing. An initial rigid transformation is randomly generated from the following intervals: the rotation along each axis in $[0°, 45°]$, and translation in $[-0.5, 0.5]$. We set up three experiments to comprehensively demonstrate the performance of DIFT.

Evaluation metrics: Following [4], [36], the performance of the comparison methods is evaluated with the root mean squared error RMSE and the mean absolute error MAE. All angular measurements are in units of degrees.

DIFT is compared against the latest approaches RIENet [38] and RegTR [31], furthermore, the baseline methods also include: ICP [35], PointNetLK (PNetLK) [3], DCP [4], DeepGMR [27], IDAM [36], Reagent [37], PointNetLK Revisited (PNetLK_R) [2], and RGM [6]. The quantitative analysis is shown in Table. I. DIFT obviously outperforms other methods in all three experiments. (1) Table I(scene 1) shows that our method achieves the best performance on clean point clouds. Compared with the second-best method, RGM, our method reduces the registration errors significantly. These results confirm the effectiveness of our PSE module in extracting structural features and the GMCCE in distinguishing correspondences. (2) We crop points to obtain a point cloud pair with an overlap rate of 60% (IoU). Then, Gaussian noise sampled from $\mathcal{N}(0, 0.01)$ and clipped to $[-0.001, 0.001]$ is added to the point clouds. Table I(scene 2) exhibits that our method outperforms all other methods, indicating that PFT enables DIFT to precisely identify common structures. (3) To evaluate the robustness against high noise, Gaussian noise independently sampled from $\mathcal{N}(0, 0.01)$ and clipped to $[-0.05, 0.05]$ is added to the point cloud pair with 60% overlap. Table I(scene 3) shows that DIFT achieves superior performance, revealing that DIFT models global relations and exhibits superior robustness against high noise.

TABLE II

PERFORMANCE ON 3DMATCH AND 3DLOMATCH BENCHMARKS. THE THREE BEST RESULTS ARE HIGHLIGHTED IN **RED**, **GREEN**, **BLUE**.

Method	Reference	3DMatch		3DLoMatch	
		RRE(°)	RTE(m)	RRE(°)	RTE(m)
3DSN [41]	CVPR 2019	2.199	0.071	3.528	0.103
FCGF [24]	CVPR 2019	2.149	0.070	3.743	0.100
D3Feat [42]	CVPR 2020	2.161	0.067	3.361	0.103
DGR [1]	CVPR 2020	2.103	0.067	3.954	0.113
PCAM [43]	ICCV 2021	**1.808**	**0.059**	3.529	0.099
OMNet [44]	ICCV 2021	4.166	0.105	7.299	0.151
Predator [45]	CVPR 2021	2.029	0.064	**3.048**	**0.093**
CoFiNet [46]	Neurips 2021	2.449	0.067	5.443	0.155
SC^2PCR [47]	CVPR 2022	2.08	0.065	3.46	0.095
Lepard [48]	CVPR 2022	**1.96**	**0.060**	**3.17**	**0.089**
Ours	-	**1.421**	**0.047**	**2.620**	**0.075**

B. Matching and Registration Performance on 3DMatch

3DMatch: To further exhibit the performance of our method in real-world registration, experiments on 3DMatch [49] are conducted. This dataset comprises 46 scenes for training and an additional 16 scenes for validation and testing. Comparison methods are evaluated on both 3DMatch [49] and 3DLoMatch [45] benchmarks with $> 30\%$ and $10 - 30\%$ overlap ratios, respectively. Following [45], the downsampled data are utilized and processed.

Evaluation metrics: Following [42], [45], all methods are evaluated with the relative rotation errors RRE (the geodesic distance between the estimated and GT rotation matrices), and relative translation errors RTE (the Euclidean distance between the estimated and GT translations).

DIFT is compared against the latest approaches Lepard [48] and SC^2PCR [47], furthermore, as well as representative methods on 3DMatch and 3DLoMatch: 3DSN [41], FCGF [24], D3Feat [42], Predator [45], CoFiNet [46], PCAM [43], DGR [1], OMNet [44]. Qualitative results are shown in Fig. 4 and quantitative comparisons are summarized in Table. II. The results show that our method aligns real-world point clouds precisely even at low overlap rates and outperforms other methods on both 3DMatch and 3DLoMatch. Com-

TABLE III

COMPUTATIONAL TIME IN SECONDS.

Methods	ICP	PNetLK	DCP_V2	DeepGMR	IDAM	PNetLK_R	Reagent	RGM	RIENet	RegTR	Ours
ModelNet40 (scene 3)	0.0022	0.022	0.014	0.022	0.017	0.073	0.021	0.119	0.033	0.029	0.045
Methods	3DSN	FCGF	D3Feat	DGR	PCAM	OMNet	Predator	CoFiNet	SC^2PCR	Lepard	Ours
3DMatch	30.234	1.562	0.916	1.741	1.786	0.012	1.572	0.414	0.080	0.522	0.115

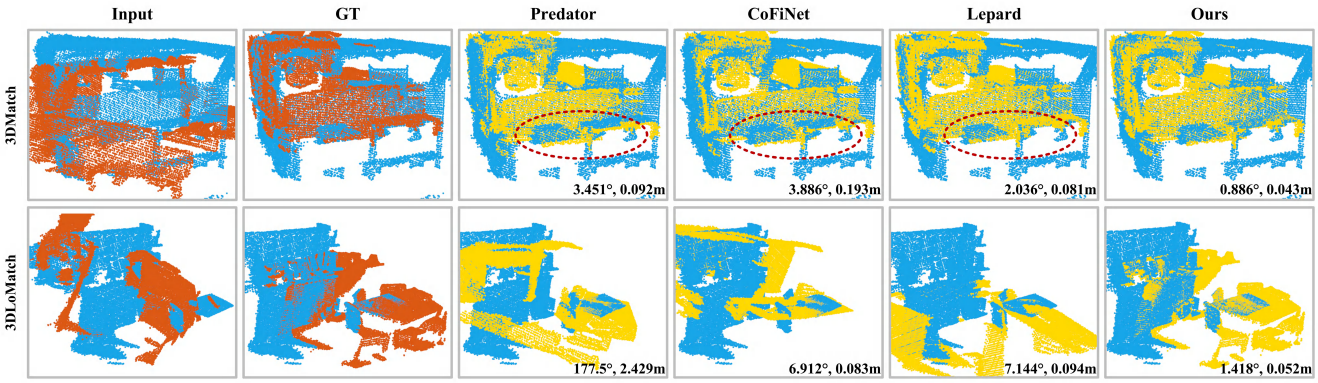

Fig. 4. Qualitative registration results on 3DMatch and 3DLoMatch. The registration errors are shown in the lower right corner.

TABLE IV

ABLATION RESULTS ON MODELNET40. SR DENOTES SUCCESS RATE.

Method	R_{RMSE}	R_{MAE}	t_{RMSE}	t_{MAE}	SR
$DIFT_{w/\ DGCNN}$	18.07	12.19	0.069	0.049	7.5%
$DIFT_{w/o\ PSE}$	41.84	28.74	0.247	0.197	1.2%
$DIFT_{w/\ SW}$	7.06	1.618	0.020	0.009	71%
$DIFT_{w/o\ PE}$	18.02	6.013	0.091	0.038	55.3%
$DIFT_{w/o\ PFT}$	20.49	7.311	0.117	0.040	54.1%
$DIFT_{w/o\ GMCCE}$	3.813	0.98	0.011	0.004	79.1%
DIFT	**0.593**	**0.286**	**0.0056**	**0.0021**	**96.4%**

pared with the second-best method for each indicator, DIFT improves the rotation and translation accuracy by 15%–25%. The experimental results clearly verify that our method accurately and robustly aligns real-world point clouds.

C. Ablation Studies

To analyze the effectiveness of the proposed three key components, ablation studies on ModelNet40 (scene3: high-noise partial) are conducted by comparing the performance of variants. The experimental results are summarized in Table IV. Specifically, a registration is counted as successful if the rotation and translation errors are less than $(1^\circ, 0.01)$.

PSE: $DIFT_{w/\ DGCNN}$ substitutes the PSE module with the DGCNN module [30], and $DIFT_{w/o\ PSE}$ is designed to exclude the PSE network, both lead to the significant drop in registration accuracy. The results demonstrate the effectiveness of the PSE network in modeling global relations and identifying structural characteristics.

PFT: $DIFT_{w/\ SW}$ utilizes a shallow-wide architecture, and $DIFT_{w/o\ PE}$ excludes the positional encoding network. The underperformance of both signifies that deep-narrow architecture facilitates the information interaction and positional encoding network enables Transformers to directly learn the position information. $DIFT_{w/o\ PFT}$ excludes the PFT, the decline in accuracy indicates that PFT identifies the common structure and extracts discriminative features.

GMCCE: $DIFT_{w/o\ GMCCE}$ is designed to exclude the GMCCE module, the success ratio drops to 79.1%, which verifies that GMCCE improves the registration accuracy with the advantage of distinguishing correspondences.

D. Efficiency Evaluation

We profile the inference time of different methods on a desktop computer with an Intel I7-10700 CPU, and an Nvidia RTX 3090 GPU. As shown in Table. III, DIFT achieves a promising balance between precise alignment and computational efficiency. Considering the significant performance in terms of accuracy, robustness to noise and low overlap rates, the computational time of DIFT is satisfactory.

V. CONCLUSION

In this work, we explore and propose a novel Transformer framework DIFT for point cloud registration. To overcome the limitations of previous Transformer-based methods, PSE is utilized to model dependencies in the entire point cloud and structurize the point cloud, enhancing the robustness to noise. Subsequently, PFT is proposed to improve the discrimination of extracted features by facilitating deep information interaction. Moreover, GMCCE is leveraged to further improve registration accuracy by detecting inliers based on geometric consistency. Extensive experiments conducted on ModelNet40 and 3DMatch exhibit the superiority of our method, and demonstrate the potential of the full Transformer framework in point cloud registration tasks.

REFERENCES

[1] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2514–2523, 2020.

[2] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Pointnetlk revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12763–12772, 2021.

[3] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7163–7172, 2019.

[4] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3523–3532, 2019.

[5] Chenghao Shi, Xieyuanli Chen, Kaihong Huang, Junhao Xiao, Huimin Lu, and Cyrill Stachniss. Keypoint matching for point cloud registration using multiplex dynamic graph attention networks. *IEEE Robotics and Automation Letters*, 2021.

[6] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8893–8902, 2021.

[7] Taewon Min, Eunseok Kim, and Inwook Shim. Geometry guided network for point cloud registration. *IEEE Robotics and Automation Letters*, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[13] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[15] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[18] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition, 2021.

[19] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. *arXiv preprint arXiv:1910.12240*, 2019.

[20] Amir Hertz, Rana Hanocka, Raja Giryes, and Daniel Cohen-Or. Pointgmm: A neural gmm network for point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12054–12063, 2020.

[21] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A. Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3412–3432, 2021.

[22] Zijin Du, Hailiang Ye, and Feilong Cao. A novel local-global graph convolutional method for point cloud semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022.

[23] Xuesong Zhang, Yan Zhuang, Huosheng Hu, and Wei Wang. 3-d laser-based multiclass and multiview object detection in cluttered indoor scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 28(1):177–190, 2017.

[24] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019.

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[26] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia, 1981.

[27] Wentao Yuan, Benjamin Eckart, Kihwan Kim, Varun Jampani, Dieter Fox, and Jan Kautz. Deepgmr: Learning latent gaussian mixture models for registration. In *European Conference on Computer Vision*, pages 733–750. Springer, 2020.

[28] Zi Jian Yew and Gim Hee Lee. Rpm-net: Robust point matching using learned features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11824–11833, 2020.

[29] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[30] Anh Viet Phan, Minh Le Nguyen, Yen Lam Hoang Nguyen, and Lam Thu Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018.

[31] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6677–6686, 2022.

[32] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[33] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[35] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[36] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. 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*, pages 378–394. Springer, 2020.

[37] Dominik Bauer, Timothy Patten, and Markus Vincze. Reagent: Point cloud registration using imitation and reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14586–14594, 2021.

[38] Yaqi Shen, Le Hui, Haobo Jiang, Jin Xie, and Jian Yang. Reliable inlier evaluation for unsupervised point cloud registration. *arXiv preprint arXiv:2202.11292*, 2022.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[41] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5545–5554, 2019.

[42] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and descrip-

tion of 3d local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6359–6367, 2020.

[43] Anh-Quan Cao, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Pcam: Product of cross-attention matrices for rigid registration of point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13229–13238, 2021.

[44] Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3132–3141, 2021.

[45] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021.

[46] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021.

[47] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13221–13231, 2022.

[48] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5554–5564, 2022.

[49] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017.