Full Transformer Framework for Robust Point Cloud Registration With Deep Information Interaction
Guangyan Chen, Graduate Student Member, IEEE, Meiling Wang, Qingxiang Zhang, Li Yuan, and Yufeng Yue, Member, IEEE

Abstract—Point cloud registration is an essential technology in computer vision and robotics. Recently, transformer-based methods have achieved advanced performance in point cloud registration by utilizing the advantages of the transformer in order-invariance and modeling dependencies to aggregate information. However, they still suffer from indistinct feature extraction, sensitivity to noise, and outliers, owing to three major limitations: 1) the adoption of CNNs fails to model global relations due to their local receptive fields, resulting in extracted features susceptible to noise; 2) the shallow-wide architecture of transformers and the lack of positional information lead to indistinct feature extraction due to inefficient information interaction; and 3) the insufficient consideration of geometrical compatibility leads to the ambiguous identification of incorrect correspondences. To address the above-mentioned limitations, a novel full transformer network for point cloud registration is proposed, named the deep interaction transformer (DIT), which incorporates: 1) a point cloud structure extractor (PSE) to retrieve structural information and model global relations with the local feature integrator (LFI) and transformer encoders; 2) a deep-narrow point feature transformer (PFT) to facilitate deep information interaction across a pair of point clouds with positional information, such that transformers establish comprehensive associations and directly learn the relative position between points; and 3) a geometric matching-based correspondence confidence evaluation (GMCCE) method to measure spatial consistency and estimate correspondence confidence by the designed triangulated descriptor. Extensive experiments on the ModelNet40, ScanObjectNN, and 3DMatch datasets demonstrate that our method is capable of precisely aligning point clouds, consequently, achieving superior performance compared with state-of-the-art methods. The code is publicly available at https://github.com/CGuangyan-BIT/DIT.

Index Terms—Deep learning, full transformer, information interaction, point cloud registration, spatial consistency.

I. INTRODUCTION

POINT cloud registration aims to calculate a rigid transformation for aligning a pair of point clouds, and it has long been a fundamental technology in computer vision and robotics, such as 3-D reconstruction and simultaneous localization and mapping (SLAM). In recent decades, point cloud registration has developed from traditional methods [1], [2], [3], [4], [5], and convolutional neural network (CNN)-based methods [6], [7], [8], to recent transformer-based methods [9], [10], [11], [12].

The most widely known traditional method is the iterative closest point (ICP) [1], which iteratively alternates between establishing correspondences and calculating a transformation. However, ICP and its variants [3], [13] are prone to falling into local minima when the initial errors are large. To increase registration accuracy, recently proposed CNN-based methods integrate CNNs to extract features and establish point-to-point correspondences based on feature similarity. However, they extract features from each point cloud separately, leading to obstacles in identifying common structures between two point clouds and extracting discriminative features.

Attention-based models, such as the transformer [14], have achieved tremendous performance in natural language processing (NLP) [15], [16], [17], [18] and computer vision tasks [19], [20], [21], [22], [23], [24]. Inspired by their success, researchers have exploited its adaptability in point cloud processing tasks, such as classification [25] and segmentation [26], and recently achieved promising results in these areas. Following the advances of the transformer in terms of order-invariance and modeling dependencies, recent point cloud registration studies [9], [11], [27], [28], [29] have investigated the incorporation of transformer models. Although point cloud registration shares many similarities with other point cloud processing tasks, there are two fundamental differences: 1) other tasks process single or multiple point clouds in a unified coordinate system, while registration is in the different coordinate systems and 2) other tasks mainly cluster points into limited categories, however, registration distinguishes each point pair and minimizes the difference within pairs. The above-summarized differences require registration methods to extract more representative and discriminative features.
Most of the recent registration methods [9], [10], [11], [29] take advantage of the attention mechanism to establish associations across two point clouds for information aggregation, enabling one point cloud to perceive another point cloud. However, substantial gaps remain in terms of modeling global relations, enhancing feature richness, and detecting inliers: 1) current methods mainly leverage 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 (SW) transformers and the lack of positional information lead to difficulty in enhancing feature richness and extracting distinct features; and 3) spatial consistency is not fully employed in inlier detection modules, which results in a large proportion of incorrect correspondences being preserved. Therefore, the above-mentioned limitations consequently reduce registration accuracy.

We observed a key similarity between NLP and point cloud registration, both of which establish associations between units representing the same content but with different expressions. Concretely, NLP establishes the connections between words with the same meaning in different languages, and point cloud registration connects the corresponding points in different coordinate systems. Motivated by this observation and the limitations of previous transformer-based methods, a full transformer framework named deep interaction transformer (DIT), which takes advantage of the transformer architecture and deep information interaction, is proposed by extending [30]. Deep information interaction establishes deep associations to facilitate information exchange, which significantly improves the discrimination of features, as visualized in Fig. 1. The proposed method is compared with extensive registration algorithms, achieving superior performance in terms of accuracy and robustness. The main contributions are fourfold:

1) A point cloud structure extractor (PSE) is proposed to integrate structural information and model global relations. In PSE, transformer encoders are adopted to model dependencies in the entire point cloud, enhancing the robustness to noise, Furthermore, the local feature integrator (LFI) is designed to structure the point cloud, which addresses the limitation of transformers in extracting structural features.

2) A point feature transformer (PFT) is proposed to increase the richness of feature representation. In PFT, deep-narrow transformers are adopted to establish deep associations, moreover, a positional encoding network is introduced to model the relative position between points.

3) A geometric matching-based correspondence confidence evaluation (GMCECE) method is proposed to estimate the correspondence confidence based on geometric constraints. Specifically, a rotation-invariant triangulated descriptor is designed to measure geometrical compatibility.

4) A systematic DIT framework that takes advantage of transformers to improve the discrimination of features, robustness to noise, and outliers, is proposed. The code for DIT is publicly available, and it outperforms SOTA methods with respect to accuracy and robustness.

II. RELATED WORK

A. Traditional Registration Methods

The most representative traditional method is the ICP algorithm [1], which iteratively alternates between finding the closest point pairs as correspondences and calculating the transformation based on established correspondences. However, ICP and its variants [3], [13] often converge to local minima when the initial position is far from the global minimum. There is a large volume of works [31], [32] that attempt to improve the robustness of ICP under poor initialization. In Gaussian mixture models (GMMs) [33], the registration problem is reformulated as the alignment of two probability distributions. However, these methods still require a warm initialization due to their nonconvex objective functions. In globally optimal ICP (Go-ICP) [2], the branch-and-bound (BnB) method is applied to search over SE(3) space to achieve global convergence, but the computational complexity is much higher than that of ICP. Fast global registration (FGR) [4] relies on optimizing a global objective function to align a pair of point clouds without updating correspondences.

In addition, handcrafted local features, such as fast point feature histograms (FPFH) [5], are also designed to establish correspondences through feature matching. However, these methods are sensitive to partially overlapping point clouds and large initial errors.

B. CNN-Based Registration Methods

The success of deep learning in point cloud processing [34], [35], [36], [37] enables its application in point cloud registration. One pioneering work is PointNetLK [8], which extracts global features using PointNet [38] and applies the inverse compositional Lucas–Kanade (IC-LK) algorithm [39] to align two point clouds. PointNetLK Revisited [7] has been 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) [40] 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) [41] is proposed to apply the Sinkhorn [42] 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 the source and target point clouds, which are inefficient in discriminative feature extraction and contextual information aggregation, especially in partial-to-partial point cloud registration tasks.

C. Transformer-Based Registration Methods

Inspired by the success of transformers in NLP and computer vision, researchers have begun to apply transformers to extract contextual information between two point clouds. Deep closest point (DCP) [9] extracts features using a dynamic graph CNN (DGCNN) [43] and utilizes a transformer [14] to aggregate information. However, DCP lacks an overall understanding of the point cloud due to its local receptive field, which results in sensitivity to noise. A multiplex dynamic graph attention network (MDGAT) [10] dynamically constructs a multiplex graph based on an attention mechanism. A geometry-guided network [12] is proposed to encode global and local features with a fully connected graph based on the self-attention mechanism. Recent registration transformer (REGTR) [44] utilizes attention layers to generate correspondences directly. Robust graph matching (RGM) method [11] adopts a transformer to aggregate information along soft graph edges. However, the edge adjacency matrices are indistinct due to the SW architecture of the transformer, which leads to limited registration accuracy. In summary, these methods mainly model local relations by directly adopting convolution encoders, which prevents them from modeling global relations. Furthermore, the information interaction in these methods is inefficient due to the SW transformer architecture and the lack of positional information.

III. PROPOSED DEEP INTERACTION TRANSFORMER

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

The overall pipeline of DIT is shown in Fig. 2. During training, the registration pipeline starts by 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 $\Phi_X$ and $\Phi_Y$. 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 $S$ between the corresponding feature vectors $\{\Phi_{x_i}, \Phi_{y_j}\}$. During testing, the GMCCE module is introduced to evaluate the correspondence confidence $\tilde{C}$, then the weighted Procrustes module estimates the optimal transformation based on the confidence $\tilde{C}$ instead of the similarity $S$.

A. Point Cloud Structure Extractor

To model dependencies in the entire point cloud and enhance the robustness to noise, the PSE module is designed. Fig. 2(a) shows the PSE module, which consists of two types of components: LFIIs and transformer encoders [14].

The limitations of transformers in structural feature extraction [23], [45] lead to difficulties in convergence and the requirement for large amounts of training data. To address this issue, LFIIs are designed to progressively structurize the point cloud. As detailed in Fig. 3, to identify the neighboring structures, the $n_{th}$ LFI ($n = 1, \ldots, N_I$) searches for the graph $G_n$ that contains the features $F_n$ of the $K$ nearest points for each point in the point cloud, where $N_I$ denotes the number of LFIIs. Specifically, the LFI applies the
limited discriminative power due to their independence from each other. Therefore, to learn the contextual information of two point clouds and extract distinct features, PFT is designed to facilitate deep information interaction. Fig. 2(b) shows the architecture of the PFT. PFT consists of a transformer-based encoder–decoder and a positional encoding network.

Since the standard transformer struggles to directly learn the relative position between points [14], a positional encoding network is introduced, which consists of fully connected layers FC, rectified linear unit ReLU activation, and sigmoid activation. The positional encoding network extracts positional information $P_X$ and $P_Y$ as

$$P_X = \text{ReLU}(\text{FC}(\text{Sigmoid}(\text{FC}(X))))$$
$$P_Y = \text{ReLU}(\text{FC}(\text{Sigmoid}(\text{FC}(Y))))$$

Subsequently, positional information $P_X$ and $P_Y$ are added to the features $F_X$ and $F_Y$ to obtain features $F_X'$ and $F_Y'$, respectively. To aggregate information from src and tgt simultaneously, a standard transformer $\phi$ is adopted, which consists of an encoder (2) and a decoder. The transformer decoder consists of a multilayer cross-attention operation (MCA), in addition to MSA (3), MLP, and LN. Taking $\phi(F_Y', F_X')$ as an example, the procedure of the decoder is defined as

$$F_X^s = \text{LN}(\text{MCA}(F_X' + F_Y'))$$
$$F_X^C = \text{LN}(\text{MCA}(F_X^s, F_X' + F_Y^s))$$
$$F_X = \text{LN}(\text{MLP}(F_X^C + F_Y^s))$$

where MCA($F_X^s$, $F_Y^s$) = MA($F_X^s$, $F_Y^s$); $F_X^s$ are obtained based on MSA, and features $F_X^C$ are acquired through the encoder, then the attention map is acquired in MCA to establish the associations between points across $X$ and $Y$, which enables $F_X^s$ to receive information from $F_Y^s$ and improve the discrimination of the extracted features.

However, due to the SW architectures utilized in previous methods, the associations established for information interaction are limited, leading to low feature richness. In this article, we instead employ a deep-narrow architecture to establish deep associations. Overall, the feature vectors $\Psi_X$ and $\Psi_Y$ generated by the transformer are formulated as

$$\Psi_X = F_X + \phi(F_Y', F_X')$$
$$\Psi_Y = F_Y + \phi(F_X', F_Y')$$

To adaptively recalibrate the channelwise features in accordance with their contribution to registration, a squeeze-and-excitation (SE) module [48] is adopted. The SE module first extracts a channel descriptor by applying average pooling to the input features, then it generates the channel weights through a neural network, and finally, rescales input features with channel weights to obtain rescaled features $\Phi_X$ and $\Phi_Y$.

In summary, by applying the positional encoding, the transformer model, and the SE module, the feature vectors $\Phi_X$ and $\Phi_Y$ generated by PFT are defined as

$$\Phi_X = \text{SE}(F_X + \phi(F_Y + P_Y, F_X + P_X))$$
$$\Phi_Y = \text{SE}(F_Y + \phi(F_X + P_X, F_Y + P_Y)).$$

Given features $\Phi_X$ and $\Phi_Y$, a set of putative correspondences $M \in \mathbb{R}^{N \times 2}$ are established by finding the most
similar features $\Phi_{xy}$ for $\Phi_{xi}$. To demonstrate the discrimination of the extracted features $\Phi_X$ and $\Phi_Y$, an intuitive example is shown in Fig. 4; in Fig. 4(a), t-SNE visualization clearly shows that the points in each pair are located in a similar area and distinguished from other points, enabling DIT to Fig. 4(b) represent overlapping regions with high similarity area and Fig. 4(c) establish accurate correspondences in overlapping regions, consequently, two point clouds are registered precisely.

C. Geometric Matching-Based Correspondence Confidence Evaluation

Since DIT is capable of extracting discriminative features and establishing accurate correspondences in the high similarity area, the extracted features $\Phi_X$ and $\Phi_Y$ are leveraged to search for the correspondence subset that is expected to have a higher inlier ratio. The subset is generated by selecting the $N$ correspondences with the highest similarity in feature space, which provides more reliable correspondences and reduces the computational time for subsequent processing. Moreover, the accuracy and robustness of registration can be further improved by filtering out incorrect correspondences in the generated subset. Therefore, the GMCCE module is designed to evaluate correspondences. As shown in Fig. 5, the descriptor employs the side length of triangles to capture geometric characteristics. The descriptor expresses the lengths and angles simultaneously, and establishes connections between sampled points, thus achieving a promising balance between the computational complexity and representational ability (Section IV-E.6).

The GMCCE module is presented in Fig. 2(c). To better illustrate the GMCCE, we detail the procedure for the evaluation of the putative correspondence $\{x_i, y_j\}$. First, KNN is adopted to search for $K$ sampled points $S_X \in \mathbb{R}^{K \times 3}$ of $x_i$ in src, then doubles $D_X \in \mathbb{R}^{K \times 2 \times 3}$ are generated by combining sampled points $S_X$ in pairs, where $K = \binom{K}{2}$. Afterward, triplets $T_X \in \mathbb{R}^{K \times 3 \times 3}$ 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 $l_X \in \mathbb{R}^{K \times 3}$ and $l_Y \in \mathbb{R}^{K \times 3}$ of the triplets $T_X$ and $T_Y$, respectively. Afterward, the overall error $E(x_i, y_j)$ is calculated by summing the $K$ smallest length errors $L$ as

$$E(x_i, y_j) = \sum \text{Mink}([L(T^1_{x}, T^2_{y}), \ldots, L(T^K_{x}, T^K_{y})])$$

$$L(T^\beta_{x}, T^\beta_{y}) = \sqrt{\sum_{i=1}^{3}(t^\beta_{x,i} - t^\beta_{y,i})^2}$$

$$+ \sum_{i=1}^{3}(t^\beta_{x,i} + t^\beta_{y,i})^2$$  \hspace{1cm} (9)

where $t^\beta_{x,i}$ denotes $i$th edge length of the triangle constructed by $T^\beta$; Mink is the operation of taking the $K$ smallest values. Finally, the confidence $\tilde{C}(x_i, y_j)$ is evaluated as

$$\tilde{C}(x_i, y_j) = \psi(2 \times \text{sigmoid}(-\lambda E(x_i, y_j)))$$  \hspace{1cm} (10)

where $\lambda$ is the parameter to adjust the sharpness of the confidence evaluation; $\psi$ is the filter to filter out correspondences with confidence values smaller than $\tau$. In summary, an integral implementation of GMCCE is presented in Algorithm 1.

D. Loss Function and Details

The overall loss function for training our DIT model consists of a transformation loss $L_t$, a cycle consistency loss $L_c$, and a discrimination loss $L_d$. By combining these terms and introducing coefficients $\alpha$ and $\beta$ to adjust the contribution of each loss term, the final loss function is constructed as

$$L = L_t + \alpha L_c + \beta L_d.$$  \hspace{1cm} (11)

1) Transformation Loss: $L_t$ measures the error between the predicted motion $R_{XY}$ and $t_{XY}$ and ground-truth motion $R^*_{XY}$ and $t^*_{XY}$ from $X$ to $Y$ as

$$L_t = \|R^T_{XY} R^*_R X_Y - I\|^2 + \|t_{XY} - t^*_{XY}\|^2.$$  \hspace{1cm} (12)

2) Cycle Consistency Loss: $L_c$ measures the consistency between the predicted motion $R_{XY}$ and $t_{XY}$ from $X$ to $Y$ and $R_{YX}$ and $t_{YX}$ from $Y$ to $X$ as

$$L_c = \|R_{XY} R_{YX} - I\|^2 + \|t_{XY} + t_{YX}\|^2.$$  \hspace{1cm} (13)
Algorithm 1 Correspondence Confidence Evaluation

**Input:** Point clouds \( X \in \mathbb{R}^{N \times 3}, Y \in \mathbb{R}^{M \times 3} \) and extracted features \( \Phi_X \in \mathbb{R}^{N \times d}, \Phi_Y \in \mathbb{R}^{M \times d} \)

**Output:** The selected correspondences \( M' \in \mathbb{R}^{N \times 2} \) and confidence \( \tilde{C} \in \mathbb{R}^{N \times 1} \)

\[ C = \text{softmax}(\text{matmul}(\Phi_X, \Phi_Y) \cdot t()) \]

# obtain correspondences \( M \) and corresponding similarity \( S \)

\[ S, I_{y \rightarrow y} \leftarrow \text{max}(C, \text{dim} = -1) \]

\[ M \leftarrow \{(i, j) \mid i \in [1, \ldots, N], j \in [1, \ldots, M] \} \]

\[ i_{dx} \leftarrow S \cdot \text{topk}(k = N, \text{dim} = -1)[1] \]

# generate correspondence subset

\[ M', X' \leftarrow M[i_{dx}], X[i_{dx}] \]

\( S_X \leftarrow \text{KNN}(X', K) \)

\( D_X \leftarrow \text{combinations}(S_X, 2) \)

# reshape and extend \( X \) for concatenate

\( E_X \leftarrow X'.\text{reshape}([N, 1, 1, 3]).\text{repeat}(1, \frac{K^2 - K}{2}, 1, 1) \)

\( T_X \leftarrow \text{concatenate}([E_X, D_X], \text{dim} = -2) \)

\( (l_T^X, l_T^Y) \leftarrow \text{GetTri}(T_X, T_Y) \)

\( \mathcal{L} \leftarrow \text{sum}(l_T^X \cdot l_T^Y)^2, \text{dim} = -1) / \text{sum}(l_T^X + l_T^Y)^2, \text{dim} = -1) \)

\( \mathcal{E} \leftarrow \text{sum}(\text{sqrt}((\text{Mink}(\mathcal{L})), \text{dim} = -1) \)

\( \tilde{C} \leftarrow \psi(2 \cdot \text{sigmoid}(-\lambda \mathcal{E}) \}

**return** \( M', \tilde{C} \)

3) **Discrimination Loss:** \( L_d \) measures the discriminative power of extracted features and the accuracy of established correspondences as

\[
L_d = -\frac{1}{\|M\|} \sum_{((x_i, y_j) \in M)} \left[ C(x_i, y_j) \times \ln S(x_i, y_j) + (1 - C(x_i, y_j)) \times \ln(1 - S(x_i, y_j)) \right]
\]

(14)

where \( C(x_i, y_j) = 1 \) if the correspondence \( \{x_i, y_j\} \) is correct; otherwise, \( C(x_i, y_j) = 0 \). \( S(x_i, y_j) \) denotes the similarity between the feature vectors \( \Phi_{xi} \) and \( \Phi_{yj} \).

4) **Implementation Details:** Each LFI layer concatenates the feature vectors from a neighborhood of \( k = 20 \) points and PSE outputs features with 64 dimensions. The MA modules in the PSE and PFT networks have \( h = 4 \) heads. In the GMCCE module, the parameters \( N = 400 \) (\( N = 200 \) on 3DMatch), \( K = 10 \), \( \lambda = 90 \), and \( \tau = 0.6 \) are obtained by grid search. For \( \alpha \) and \( \beta \) in the loss function, \( \alpha \) is set to 0.1, and \( \beta \) is set to 0.5. In addition, DIT is trained with two batch sizes over 80 epochs, using Adam [49] with an initial learning rate (LR) of \( 3^{-5} \) and a weight decay of \( 10^{-6} \). Guided by our empirical results, the multistep LR schedule is utilized, which drops the LR by 0.1 at epochs [24], [48], [64].

IV. EXPERIMENTAL RESULTS

A. **Registration Performance on ModelNet40**

1) **ModelNet40:** The ModelNet40 [53] dataset includes 12,311 meshed computer-aided design (CAD) models with 40 categories, of which 80% are designated for training and the remaining are designated for testing. We uniformly sample 1024 points as src and rescale the point cloud to a unit sphere. An initial rigid transformation is randomly generated from the following intervals: the rotation along each axis in [0°, 45°], and the translation over [−0.5, 0.5]. This initial transformation is then applied to src to obtain tgt. To comprehensively demonstrate the performance of DIT, three experiments are conducted to investigate the sharpness of the extracted features, the robustness to outliers, and noise.

2) **Comparison Methods:** DIT is compared with the latest approaches RENet [52] and RegTR [44]; furthermore, the baseline methods also include traditional algorithms: ICP [1], FGR [4], and FPFH [5] + RANSAC [54]; and recent learning-based methods: PointNetLK (PNetLK) [8], DCP [9], DeepGMR [40], IDAM [50], Reagent [51], PointNetLK revisited (PNetLK_R) [7], and RGM [11]. All experiments are evaluated on an Intel i7-10700 CPU with an RTX 3090 graphics card. ICP, FGR, and FPFH are implemented with the Intel Open3d library [55]. For the other methods, we reproduced the open-source code provided by the published papers with the provided settings and hyperparameters.

3) **Evaluation Metrics:** Following [9] and [50], the performance of each method is evaluated using the root-mean-squared error (RMSE) and the mean absolute error (MAE). All angular measurements are in units of degrees.

4) **Clean Point Clouds:** We first evaluate the performance on clean point clouds. The qualitative results are shown in Fig. 6(a) and quantitative comparisons are summarized in Table I (scene 1), our method achieves the best performance. Compared with the second-best method, RGM, our method significantly reduces the rotation and translation errors. Furthermore, the underperformance of RegTR indicates that it lacks precise alignment capabilities and verifies the ability of our PSE module to identify the structure of point clouds. The experimental results demonstrate that the PSE enhances the ability to model local structures and the deep-narrow architecture of the PFT sharpens the mapping for alignment.

5) **Low Noise Partial-to-Partial Point Clouds:** Partial-to-partial registration is much more challenging due to the existence of outliers and the difficulty in extracting contextual information. Following a similar operation of generating partial-to-partial point clouds in PRNet [29], we crop 200 points from each src and tgt to obtain a point cloud pair with an overlap rate of approximately 60% [intersection over union (IoU)]. Then, Gaussian noise sampled from \( \mathcal{N}(0, 0.01) \) and clipped to [−0.001, 0.001] is added to each point. The results of low noise partial-to-partial registration are shown in Fig. 6(b) and Table I (scene 2). Our method clearly outperforms the other methods; specifically, the rotation and translation errors are obviously reduced even compared with RENet and RGM. Furthermore, due to the lack of information aggregation, the accuracy of DeepGMR and PNetLK-R is much lower than that on clean point clouds. In addition, RegTR still struggles to accurately align point clouds. The experimental results verify the effectiveness of aggregating information across point clouds and show that DIT extracts...
contextual information by means of deep information interaction, enabling DIT to precisely identify common structures.

6) High Noise Partial-to-Partial Point Clouds: To evaluate the robustness against high noise in partial-to-partial registration tasks, similar to the operation in PRNet [29], Gaussian noise independently sampled from $\mathcal{N}(0, 0.01)$ and clipped to $[-0.05, 0.05]$ is added to each point. The other experimental settings are the same as in the low-noise experiment. The results of the high noise partial-to-partial registration are shown in Fig. 6(c) and Table I (scene 3). Our method still outperforms the other methods; specifically, our method enhances the rotation and translation accuracy by 32%–65% compared with RGM, RIENet, and RegTR. Note that the accuracy of DCP is reduced by approximately 60% compared with its accuracy in the low noise case. The results reveal that DIT is able to model global relations, thereby achieving superior robustness against high noise.

B. Accuracy and Generalization Analysis on ModelNet40

To compare the accuracy and generalization, we present the success ratio as a function of the rotation and translation thresholds ($R_{\text{thres}}$ and $t_{\text{thres}}$) in Fig. 7, and the box plots of the rotation and translation errors ($R_{\text{error}}$ and $t_{\text{error}}$) in Fig. 8, where the success ratio is the proportion of successful alignments, defined as having the errors less than the thresholds. It is clear that our method achieves the highest success ratios in most cases, indicating that DIT is capable of aligning point clouds with a high success rate despite strict accuracy requirements. Furthermore, the box plots show that DIT has a superior error distribution among the compared methods. The experimental results further demonstrate that our method achieves the highest accuracy and generalization.

1) Clean Point Clouds: Fig. 7(a) shows the success ratios on clean point clouds. Our method achieves the fastest convergence and a 100% success rate in both rotation and translation. Due to the strict convergence thresholds, several methods, such as RegTR, did not converge, indicating unsatisfactory performance for precise alignment. Only RGM and PNetLK_R are competitive with our method, but they have higher mean rotation and translation errors. Specifically, DIT achieves 92% at a rotational threshold of $3e^{-6}$ and 98% at a translation threshold of $3e^{-8}$, which are better than the 47% and 37% achieved by RGM. The results indicate that the PSE module enables DIT to identify structural features, and PFT conducts information interaction, significantly improving accuracy.

2) Low Noise Partial-to-Partial Point Clouds: As shown in Fig. 7(b), our method is the only approach with an ultimate success ratio of 99% in both rotation and translation. Furthermore, Fig. 8(b) shows that the errors of DIT are concentrated in the low error range. Most of the other methods have
success ratios of less than 85%. Only RIENet and RGM are comparable to our method, but they have higher matching errors. Concretely, the maximum rotation and translation errors of DIT are only 0.1° and 2e-4 respectively, while those of RGM reach 15° and 0.13, and those of RIENet reach 20° and 0.15. The results reveal that the deep-narrow architecture and positional encoding enable DIT to extract discriminative features and align point clouds precisely.

3) High Noise Partial-to-Partial Point Clouds: Fig. 7(c) shows the success ratios in the more challenging settings. Our method still surpasses the other methods, achieving a 98% success rate in both rotation and translation. As shown in Fig. 8(c), only RIENet, RegTR, and RGM achieve comparable performance. However, these methods yield significant errors, resulting in high mean errors in rotation and translation. The results demonstrate that modeling global relations is able to strengthen the robustness to high noise, and our full transformer framework achieves high accuracy and generalization in various registration tasks.

C. Registration Performance on ScanObjectNN

1) ScanObjectNN: To evaluate the generalization performance from synthetic to real-world data, we introduce point
clouds with segmented objects from the ScanObjectNN [56] dataset, where 2902 3-D objects are extracted. The same experimental settings as the previous experiments in Section IV-A are applied for a fair comparison.

To better reflect the generalization ability to real-world datasets, the model trained on ModelNet40 is transferred to ScanObjectNN for testing without any retraining or fine-tuning. The qualitative results and quantitative analyses are shown in Fig. 9 and Table II. DIT precisely aligns point clouds and obviously outperforms the other methods in all three experiments, which shows that our method accurately and robustly aligns point clouds not only on synthetic data but also on real-world data.

To intuitively compare the accuracy and robustness of each method on the real-world dataset, success ratios and box plots are shown in Figs. 10 and 11, respectively. DIT has higher success ratios and superior error distribution. Fig. 10(a) shows that our method significantly surpasses the other methods on clean point clouds. DIT is the only approach that ultimately achieves a 100% success rate in both rotation and translation, while RGM only reaches a 93% success rate in both, PNetLK_K reaches 87% and 68% success rates, which are significantly lower than the performance on ModelNet40. Fig. 10(b) exhibits the performance of our method on the low noise partial-to-partial point clouds. DIT achieves a 90% success rate at a rotational threshold of $5e - 4$, compared with 67% and 72% for RGM, 65% and 70% for RIENet. Furthermore, Fig. 11(b) shows that the maximum rotation and translation errors of DIT are only $1^\circ$ and 0.005, while those of RGM and RegTR exceed $10^\circ$ and 0.1. The success rates on high noise partial-to-partial point clouds clearly depict that DIT outperforms RIENet and RegTR at most thresholds. Moreover, DIT ultimately achieves a 96% success rate in rotation and translation, while RIENet only achieves 87% and 75%. In general, DIT achieves superior performance in all three different settings, and generalizing from synthetic data to real-world data has only a small impact on the accuracy of DIT; however, the accuracy of other methods evidently decreases. The experimental results demonstrate that DIT is capable of learning and extracting features with strong generalization.

D. Registration Performance on 3DMatch

1) 3DMatch: To further exhibit the performance of our method in the real-world point cloud registration, experiments on 3DMatch [57] are conducted. The 3DMatch is a real-world pairwise registration dataset that utilizes scanned point clouds of different frames as the source and target point clouds, respectively. The 3DMatch dataset contains 62 scenes, of which 46 scenes are designed for training and the remaining 16 scenes are evenly allocated for validation and testing. Comparison methods are evaluated on both 3DMatch...
Fig. 10. Success rates for varying rotation or translation error thresholds on the ScanObjectNN dataset. A registration is considered successful if the rotation (first row) or translation (second row) error is less than the thresholds (x-axis). (a) Clean point clouds. (b) Low noise partial point clouds. (c) High noise partial point clouds.

Fig. 11. Box plots of the rotation and translation errors of the ten compared methods on the ScanObjectNN dataset with three experimental settings. (a) Clean point clouds. (b) Low noise partial point clouds. (c) High noise partial point clouds.

(>30% overlap ratios) [57] and 3DLoMatch (10%–30% overlap ratios) [58] benchmarks. Following Predator [58], the voxel-grid downsampled data are utilized and processed for input.

2) Comparison Methods: DIT is compared with the latest approaches RegTR [44] (RIENet did not perform experiments on 3DMatch); furthermore, the comparison methods also include the representative methods on 3DMatch: 3DSN [59], FCGF [37], D3Feat [60], CG-SAC [61], Predator [58], CoFiNet [62], PCAM [63], DGR [6], OMNet [64], and DHVR [65]. The number of interest points based on RANSAC in the correspondence-based methods is set to 5000 (the maximum value) for the best performance. We also present the Predator$k using 1k interest points for the highest registration recall.

3) Evaluation Metrics: Following [44], the performance of each method is evaluated using the registration recall (RR) (the percentage of successful alignments, defined as having a correspondence RMSE below 0.2), 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).

The qualitative results are shown in Fig. 12 and quantitative comparisons are summarized in Table III. The results show that our method aligns a pair of real-world point clouds precisely even at low overlap rates and outperforms other methods on both the 3DMatch and 3DLoMatch benchmarks. Compared with the second-best method on the 3DMatch benchmark, RegTR, DIT improves the registration recall to 92.7% and enhances the rotation accuracy by 10%. Compared with CoFiNet on the 3DLoMatch, DIT achieves superior registration recall and reduces both RRE and RTE by 51%. Compared with RegTR on the 3DLoMatch, DIT outperforms RegTR in terms of registration recall by 3.9%. The experimental results verify that our method aligns real-world point clouds accurately.

E. Ablation Studies

1) Effect of the Proposed Components: To analyze the effectiveness of the proposed three key components (PSE, PFT, and GMCCE), ablation studies on ModelNet40 and 3DMatch are presented. Specifically, experiments on ModelNet40 are performed under high-noise partial conditions. The results on 3DMatch and 3DLoMatch benchmarks are shown in Table IV, and the results on ModelNet40 are shown in Table V and Fig. 13.

2) PSE: 1) To demonstrate the inductive bias introduced by performing KNN in geometric space, DIT_{w/FKNN} performs KNN in the feature space, DIT_{w/HKNN} performs KNN in geometric space for the first four layers and in feature space for the last four layers, Table V depicts that the success ratios drop to
that the success ratio drops to 7.5%. 3) DIT w/ ModelNet40 and 3DMatch, which verifies that GMCCE significantly improves registration accuracy with the advantage of distinguishing between inliers and outliers. 2) DCP w/GMCCE and PNetLK R_{w/GMCCE} are designed to employ GMCCE (Table V), which improves the success ratios by 43.8% and 36% on ModelNet40. The performance of these variants shows that GMCCE can also improve other methods.

5) Effect of λ and τ in GMCCE: Due to the importance of (10) in GMCCE, the sensitivity of λ and τ is further analyzed on high noise partial-to-partial point clouds of ModelNet40. As shown in Table VI, either setting λ or τ too large or too small would lead to a performance decline, as more correct correspondences are filtered out or more incorrect correspondences are preserved. Furthermore, λ = 90 and τ = 0.6 exhibit satisfying success ratios and robustness.

6) Effect of Geometric Shapes in GMCCE: The geometric shapes of the descriptor also affect the GMCCE. To explore the appropriate choice, experiments on ModelNet40 are conducted. As shown in Fig. 14(a), success rates rise as the shape becomes more complicated. The triangle exhibits a tradeoff between the success ratio and computational expense. As visualized in Fig. 14(b), the triangle achieves a superior success ratio, indicating that complex shapes are more sensitive to high noise. The experimental results indicate that the triangle shape is the appropriate choice in terms of the representative ability, computational expense, and robustness to noise.

F. Efficiency Evaluation

Real-time performance is an important indicator of point cloud registration methods. The inference time of each method including preprocessing is profiled on a desktop computer with an Intel I7-10700 CPU, and an NVIDIA RTX 3090 GPU. Specifically, ICP, FGR, and FPFH are tested on the CPU, and other methods are tested on the GPU. For ModelNet40 and ScanObjectNN, the computational time is computed by the RANSAC process, RGM takes three times as long as the consuming RANSAC process, RGM takes three times as long as the consuming RANSAC process.
as our method due to the Sinkhorn [42] outlier rejection. Considering the significant performance in terms of accuracy, robustness to noise, and low overlap rates, the computational time of DIT is satisfactory.

V. CONCLUSION

In this work, we explore and propose a novel full transformer DIT framework for point cloud registration. The DIT effectively models global relations, enhances feature richness, and distinguishes correspondences, overcoming the limitations of previous transformer-based methods. In DIT, PSE is utilized to model global relations and identify the characteristics of neighboring structures, ultimately enhancing the robustness to noise. Furthermore, PFT is proposed to improve the discrimination of extracted features by facilitating deep information interaction. Moreover, GMCCE is leveraged to improve registration accuracy by detecting inliers based on geometric consistency. Extensive experiments are conducted on ModelNet40, ScanObjectNN, and 3DMatch, exhibiting superior performance in terms of accuracy, generalization, and robustness, demonstrating the potential of the full transformer framework in point cloud registration tasks.

REFERENCES

[1] P. J. Besl and N. D. McKay, “Method for registration of 3-D shapes,” *Proc. SPIE*, vol. 1611, pp. 586–606, Apr. 1992.
[2] J. Yang, H. Li, D. Campbell, and Y. Jia, “Go-ICP: A globally optimal solution to 3D ICP point-set registration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2241–2254, Nov. 2016.
[3] A. V. Segal, D. Harchen, and S. Thrun, “Generalized-ICP” in *Robotics: Science and Systems*, vol. 2, no. 4, Seattle, WA, USA: MIT Press, Jun. 2009, p. 435.
[4] Q.-Y. Zhou, J. Park, and V. Koltun, “Fast global registration,” in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 766–782.
[5] R. B. Rusu, N. Blodow, and M. Beetz, “Fast point feature histograms (FPFH) for 3D registration,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 3212–3217.
[6] C. Choy, W. Dong, and V. Koltun, “Deep global registration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2514–2523.
[7] X. Li, J. K. Pontes, and S. Lucey, “PointNetLK revisited,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12763–12772.
[8] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, “PointNetLK: Robust & efficient point cloud registration using PointNet,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7163–7172.
[9] Y. Wang and J. Solomon, “Deep closest point: Learning representations for point cloud registration,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3523–3532.
[10] C. Shi, X. Chen, K. Huang, J. Xiao, H. Lu, and C. Stachniss, “Key-point matching for point cloud registration using multiplex dynamic graph attention networks,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8221–8228, Oct. 2021.
[11] K. Fu, S. Liu, X. Luo, and M. Wang, “Robust point cloud registration framework based on deep graph matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8893–8902.
[12] T. Min, E. Kim, and I. Shim, “Geometry guided network for point cloud registration,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7270–7277, Oct. 2021.
[13] S. Rusinkiewicz and M. Levoy, “Efficient variants of the ICP algorithm,” in *Proc. 3rd Int. Conf. 3-D Digit. Imag. Model.*, 2001, pp. 145–152.
[14] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.
[16] H. Fei, Y. Zhang, Y. Ren, and D. Ji, “Optimizing attention for sequence modeling via reinforcement learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 1–10, Feb. 2021.
[17] J. Sun, S. Wang, J. Zhang, and C. Zong, “Neural encoding and decoding with distributed sentence representations,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 589–603, Feb. 2021.
[18] K. Zhang, G. Lv, L. Wu, E. Chen, Q. Liu, and M. Wang, “LadRa-Net: Locally aware dynamic reread attention for sentence semantic matching,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 1–14, Aug. 2021.
[19] K. Jiang et al., “Multi-scale hybrid fusion network for single image deraining,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 24, 2021, doi: 10.1109/TNNLS.2021.3112235.
[20] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[21] Z. Shao, J. Han, D. Marnerides, and K. Debattista, “Region-obje ction-relation-aware dense captioning via transformer,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2022, doi: 10.1109/TNNLS.2022.3152990.
[22] P. Xu, K. Chaitanya Joshi, and X. Bresson, “Multigraph transformer for free-hand sketch recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 1–12, Apr. 2021.
[23] L. Yuan et al., “Tokens-to-Token VT: Training vision transformers from scratch on ImageNet,” 2021, arXiv:2101.11986.
[24] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, “VQLO: Vision looker for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6575–6586, Sep. 2021.
[25] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, “PCT: Point cloud transformer,” *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, 2021.
[26] J. Wang, R. Chakraborty, and S. X. Yu, “Spatial transformer for 3D point clouds,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[27] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, “A vision transformer model for convolution-free multi-label classification of satellite imagery in deforestation monitoring,” 2021, arXiv:2105.00057.
[28] Y. Wang and J. M. Solomon, “PRNet: Self-supervised learning for partial-to-partial registration,” 2019, arXiv:1910.12240.
G. Chen, M. Wang, Q. Zhang, L. Yuan, T. Liu, and Y. Yue, “Deep interactive full transformer framework for point cloud registration,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Jun. 2023, pp. 1–23.

B. Eckart, K. Kim, and J. Kautz, “HGMR: Hierarchical Gaussian mixtures for adaptive 3D registration,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 705–721.

D. Campbell and L. Peterson, “GOGM: Globally-optimal Gaussian mixture alignment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5685–5694.

B. Jian and B. C. Vemuri, “Robust point set registration using Gaussian mixture models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1633–1645, Aug. 2011.

Y. Li et al., “Deep learning for LiDAR point clouds in autonomous driving: A review,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 8, pp. 3412–3432, Aug. 2022.

Z. Du, H. Ye, and F. Cao, “A novel local-global graph convolutional method for point cloud semantic segmentation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3155282.

X. Zhang, Y. Zhuang, H. Hu, and W. Wang, “3-D laser-based multiclass and multiview object detection in cluttered indoor scenes,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 1, pp. 177–190, Jan. 2017.

C. Choy, J. Park, and V. Koltun, “Fully convolutional geometric features,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8958–8966.

R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 652–660.

B. D. Lucas et al., “An iterative image registration technique with an application to stereo vision,” in Proc. 7th Int.Joint Conf. Artif. Intell., Vancouver, BC, Canada, 1981, pp. 674–679.

W. Yuan, B. Eckart, K. Kim, V. Jampani, D. Fox, and J. Kautz, “DeepGMRF: Learning latent Gaussian mixture models for registration,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 733–750.

Z. J. Yew and G. H. Lee, “RPM-Net: Robust point matching using learned features,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11824–11833.

R. Sinkhorn, “A relationship between arbitrary positive matrices and doubly stochastic matrices,” Ann. Math. Statist., vol. 35, no. 2, pp. 876–879, Jun. 1964.

A. V. Phan, M. L. Nguyen, Y. L. H. Nguyen, and L. T. Bui, “DGCNN: A convolutional neural network over large-scale labeled graphs,” Neural Netw., vol. 108, pp. 533–543, Dec. 2018.

Z. J. Yew and G. H. Lee, “REGTR: End-to-end point cloud correspondences with transformers,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 6677–6686.

A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, “Point transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 16759–16768.

L. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 1–27, 2008.

J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

J. Li, C. Zhang, Z. Xu, H. Zhou, and C. Zhang, “Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration,” in Proc. 16th Eur. Conf. Comput. Vis. Glasgow, U.K.: Springer, Aug. 2020, pp. 378–394.

D. Bauer, T. Pattan, and M. Vincze, “ReAgent: Point cloud registration using imitation and reinforcement learning,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14586–14594.

Y. Shen, L. Hui, H. Jiang, J. Xie, and J. Yang, “Reliable inlier evaluation for unsupervised point cloud registration,” 2022, arXiv:2202.11292.

Z. Wu et al., “3D ShapeNets: A deep representation for volumetric shapes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1912–1920.

M. A. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.

Q.-Y. Zhou, J. Park, and V. Koltun, “Open3D: A modern library for 3D data processing,” 2018, arXiv:1801.09847.

M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, “Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1588–1597.

A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, “3DMatch: Learning local geometric descriptors from RGB-D reconstructions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1802–1811.

S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, “PREDATOR: Registration of 3D point clouds with low overlap,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4267–4276.

Z. Gojcic, C. Zhou, J. D. Wegner, and A. Wieser, “The perfect match: 3D point cloud matching with smoothed densities,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5545–5554.

X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, “D3Feat: Joint learning of dense detection and description of 3D local features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6359–6367.

S. Quan and J. Yang, “Compatibility-guided sampling consensus for 3-D point cloud registration,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 10, pp. 7380–7392, Oct. 2020.

H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic, “CoFiNet: Reliable coarse-to-fine correspondences for robust point cloud registration,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 23872–23884.

A.-Q. Cao, G. Puy, A. Boulch, and R. Marlet, “PCAM: Product of cross-attention matrices for rigid registration of point clouds,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 13229–13238.

X. Xu, S. Liu, G. Wang, G. Liu, and B. Zeng, “OMNet: Learning overlapping mask for partial-to-partial point cloud registration,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 3132–3141.

J. Lee, S. Kim, M. Cho, and J. Park, “Deep Hough voting for robust global registration,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15994–16003.

Guangyang Chen (Graduate Student Member, IEEE) received the B.S. degree in automation from the Beijing Institute of Technology, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree in intelligent navigation with the School of Automation. His research interests include 3-D deep learning, machine learning, and collaborative localization for robot systems.

Meiling Wang received the B.S. and M.S. degrees in automation and the Ph.D. degree in navigation, guidance, and control from the Beijing Institute of Technology, Beijing, China, in 1992, 1995, and 2007, respectively. She was a Visiting Scholar with the University of California at San Diego, San Diego, CA, USA, in 2004. Since 1995, she has been with the Beijing Institute of Technology, where she is currently a Professor and the Director of the Integrated Navigation and Intelligent Navigation Laboratory. Her research interests include advanced technology of sensing and detecting and vehicle intelligent navigation.
Qingxiang Zhang received the B.S. degree in electrical engineering and automation from the Beijing Institute of Technology, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree in intelligent navigation with the School of Automation.

His research interests include semantic understanding and collaborative localization for robot systems.

Li Yuan received the B.Eng. degree from the University of Science and Technology of China, Hefei, China, in 2017, and the Ph.D. degree from the National University of Singapore, Singapore, in 2021.

He is currently a tenure-track Assistant Professor with the School of Electrical and Computer Engineering, Peking University, the Peng Cheng Laboratory, Shenzhen, China. He has authored or coauthored more than 20 journals/conference papers in computer vision and machine learning. His research interests include computer vision and deep learning.

Yufeng Yue (Member, IEEE) received the B.Eng. degree in automation from the Beijing Institute of Technology, Beijing, China, in 2014, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2019.

He is currently a Professor with the School of Automation, Beijing Institute of Technology. He has authored or coauthored a book in Springer and more than 40 journals/conference papers. His research interests include perception, mapping, and navigation for autonomous robotics.