

RDMNet: Reliable Dense Matching Based Point Cloud Registration for Autonomous Driving

Chenghao Shi, Xie Yuanli Chen, Huimin Lu*, Wenbang Deng, Junhao Xiao*, Bin Dai

arXiv:2303.18084v1 [cs.CV] 31 Mar 2023

Abstract—Point cloud registration is an important task in robotics and autonomous driving to estimate the ego-motion of the vehicle. Recent advances following the coarse-to-fine manner show promising potential in point cloud registration. However, existing methods rely on good superpoint correspondences, which are hard to be obtained reliably and efficiently, thus resulting in less robust and accurate point cloud registration. In this paper, we propose a novel network, named RDMNet, to find dense point correspondences coarse-to-fine and improve final pose estimation based on such reliable correspondences. Our RDMNet uses a devised 3D-RoFormer mechanism to first extract distinctive superpoints and generates reliable superpoints matches between two point clouds. The proposed 3D-RoFormer fuses 3D position information into the transformer network, efficiently exploiting point clouds’ contextual and geometric information to generate robust superpoint correspondences. RDMNet then propagates the sparse superpoints matches to dense point matches using the neighborhood information for accurate point cloud registration. We extensively evaluate our method on multiple datasets from different environments. The experimental results demonstrate that our method outperforms existing state-of-the-art approaches in all tested datasets with a strong generalization ability.

Index Terms—Autonomous Driving, 3D Registration, Deep Learning, Point Cloud Data Processing

I. INTRODUCTION

Point cloud registration is a fundamental problem in computer vision, robotics, and autonomous driving. It aims to estimate the transformation between pairs of partially overlapped point clouds. The correspondence-based methods [1]–[3] are the current domination. They first find the data association, such as point matches between two LiDAR point clouds. Based on that, they then compute the relative transformation straightforwardly with a singular value decomposition (SVD) or a robust estimator, e.g., RANSAC [4]. To balance the computation consumption and correspondence quality, most existing methods find the association on the downsampled sparse points or keypoints [1], [3], [5]. However, downsampling will inevitably make part of the points lose their corresponding points, which degrades the registration performance.

Inspired by works in image matching, recent advances [2], [7] utilize the coarse-to-fine mechanism show remarkable

Chenghao Shi, Xie Yuanli Chen, Huimin Lu, Wenbang Deng and Junhao Xiao are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China.

Bin Dai is with the Unmanned Systems Research Center, National Innovation Institution of Defense Technology, Beijing, China.

Corresponding author: Junhao Xiao, Huimin Lu, e-mail: junhao.xiao@ieee.org, lhmnew@nudt.edu.cn.

This work was supported in part by the National Science Foundation of China under Grant U1913202, U22A2059 and U1813205, as well as Major Project of Natural Science Foundation of Hunan Province under Grant 2021JC0004.

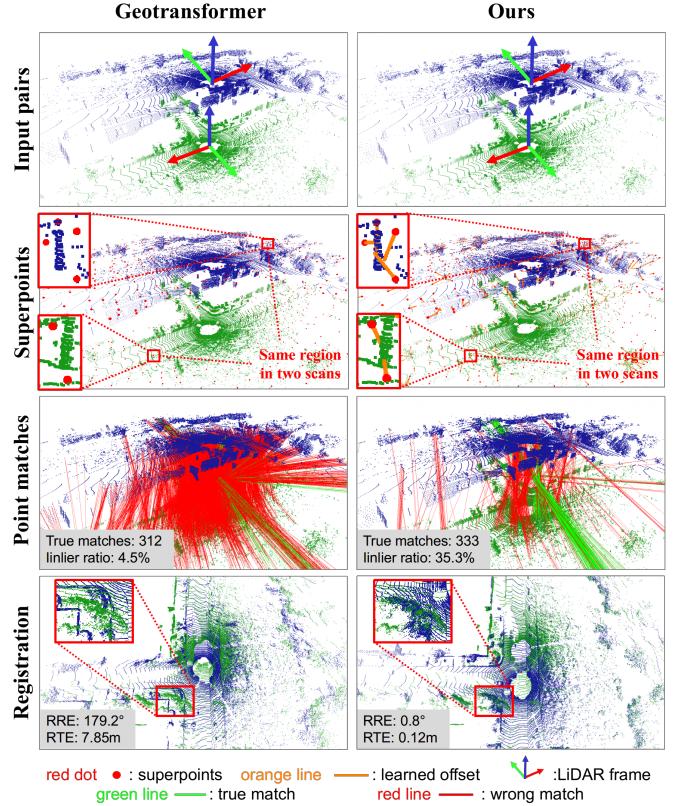


Fig. 1: Point cloud registration under a challenging situation. We compare our method (right column) against GeoTransformer [6] (left column). Given two point clouds (first row), we first extract the sampled points uniformly distributed in the point cloud. Geotransformer uses them directly as superpoints in (second left), while our RDMNet adds the learned offsets to the sampled points and generates better superpoints near the geometrically significant regions (second right). The orange lines show the learned offsets, which make superpoints from both point clouds fall closer in the geometrically significant regions. Using our proposed 3D-RoFormer to generate high-quality superpoint correspondences, our RDMNet subsequently finds better dense-point correspondences (third right) compared to the baseline method (third left). In the end, our method successfully registers the two scans (bottom right) while the baseline method fails (bottom left).

performance on point cloud registration. The coarse-to-fine mechanism [2], [7] downsamples the point cloud into sparse superpoints and uses these superpoints to split the point cloud into point patches. On the coarse level, it finds the superpoint (patch) correspondences depending on the overlapped area of two patches. On the fine level, the superpoint correspondences are then propagated to dense point matches based on neighborhood consensus, i.e., finding the point matches only from the matched patches. This limits the search space to a reasonable range, which not only reduces the computational complexity,

but also makes the established matches more reliable. However, the performance of the final matches heavily relies on the quality of the superpoint correspondences.

In this paper, we dig into the properties of the superpoint that affect the final performance. As a powerful contextual information encoder, transformer [8] has been adapted to multiple point cloud learning tasks [9]. Existing methods [2], [6], [7] exploit transformers to increase the robustness of superpoint matches. However, the vanilla transformer used in CoFiNet [2] lacks geometric information, which hinders the performance of the position-sensitive point cloud registration. GeoTransformer [6] infuses the pair-wise distance and triplet-wise angular information into the transformer, while Neg-Net [7] constructs point pair features based on the geometric information [10]. Although yielding promising results, they are computationally expensive and neglect the distribution of the superpoints, thus leading to suboptimal point-matching results.

To tackle the above-mentioned problems, we propose the RDMNet to exploit both the contextual and geometric information of the point cloud and generate reliable dense-point correspondence for point cloud registration. The core technique in RDMNet is our devised novel transformation-invariant attention mechanism, named 3D-RoFormer. It encodes the 3D position into a deep rotation matrix and naturally incorporates explicit relative position dependency into the self-attention calculation, thus becoming transformation-invariant while keeping lightweight and fast. For the superpoint distribution, RDMNet uses a superpoint detection module to first uniformly sample points over the whole point cloud and then learn the offset for each point, making the superpoint pairs more compact and falling in significant regions. Fig. 1 demonstrates that our RDMNet extracts more compact superpoint pairs and finds more reliable dense-point correspondences compared to the baseline methods.

To thoroughly evaluate our method, we conduct experiments on multiple outdoor datasets, including KITTI [11], KITTI-360 [12], Apollo [13], Mulran [14], and a self-recorded dataset with our own mobile robot in a campus environment. Note that we only train our method on the training data of the KITTI dataset and directly apply it to other datasets collected by different LiDAR sensors in different environments. The experimental results show that our method outperforms the state-of-the-art methods in terms of both superpoint matching and pose estimation with strong generalization ability.

To sum up, our main contributions are:

- A novel transformer, 3D-RoFormer, that efficiently learns the contextual and geometric features for superpoint matching with limited computation and storage cost.
- A novel network RDMNet that generates reliable superpoints and dense-point correspondences to achieve state-of-the-art point cloud registration performance.
- Extensive evaluations on multiple outdoor datasets while only trained on the KITTI dataset show that our RDMNet achieves superior performance with strong generalization ability compared to other state-of-the-art methods.

We will make the implementation of our method open-source.

II. RELATED WORK

Point cloud registration refers to finding the relative spatial transformation that aligns two point clouds. The existing methods can be broadly classified into two categories: correspondence-free and correspondence-based.

The correspondence-free methods transform the registration problem into a regression problem. Early work like PointNetLK proposed by Aoki *et al.* [15] first extracts the feature of the point cloud using PointNet [16] and then regresses the transformation from the features. Zheng *et al.* [17] utilize a similar idea with PointNetLK and further refine the result in an iterative computation architecture. Other works such as the one by Huang *et al.* [18] solve the registration problem by minimizing the feature-metric projection error. Such methods struggle to construct reliable regression models, and registration accuracy is not guaranteed.

The correspondence-based methods first extract correspondences between two point clouds and then compute the transformation using a direct solver or a robust estimator. Extracting correct correspondences is the most challenging part of such methods. The standard correspondence-based approach is the iterative closest point (ICP) algorithm [19] and its numerous variants [20], [21]. They find correspondences using the nearest neighbor search or other heuristics iteratively, thus heavily relying on good initial estimation for transformation. Recent work [9] follows the idea of ICP and establishes the soft correspondences in the learned feature space, which relaxes the requirement of good initial guesses. However, the computational complexity and global searching limit the application of these methods to large-scale point clouds.

Different from the above ICP-like methods, keypoint-based methods find correspondences on sparse points generated by either uniformly sampling [22]–[24] or keypoint detection [3], [5], [25]. PPFNet and PPF-FoldNet proposed by Deng *et al.* [22], [23] introduce point pair feature (PPF) combined with PointNet to produce local patch representation for matching. Different from PPFNet and PPF-FoldNet establishing correspondences on uniform sampling points, keypoint-based methods sample points according to pre-defined [26] or learned saliency [3], [5], [25] for higher repeatability. In order to handle the low overlap situation, PREDATOR proposed by Huang *et al.* [1] extracts the points that are not only salient but also lie in the overlap region. Based on that, Zhu *et al.* [6] recently proposed NgeNet to augment the features with point pair information and use a multi-level consistency voting to improve discrimination.

Inspired by recent advances in image matching, CoFiNet proposed by Yu *et al.* [2] and GeoTransformer proposed by Qin *et al.* [7] utilize a coarse-to-fine mechanism that first finds reliable sparse point patch correspondences and then propagates the sparse correspondences to dense point matches. Benefiting from the highly efficient backbone, they are able to obtain dense descriptors for large point clouds. The coarse-to-fine mechanism also greatly reduces the search space and increases the matching reliability. HRegNet proposed by Lu *et al.* [27] also utilizes the coarse-to-fine scheme, but different

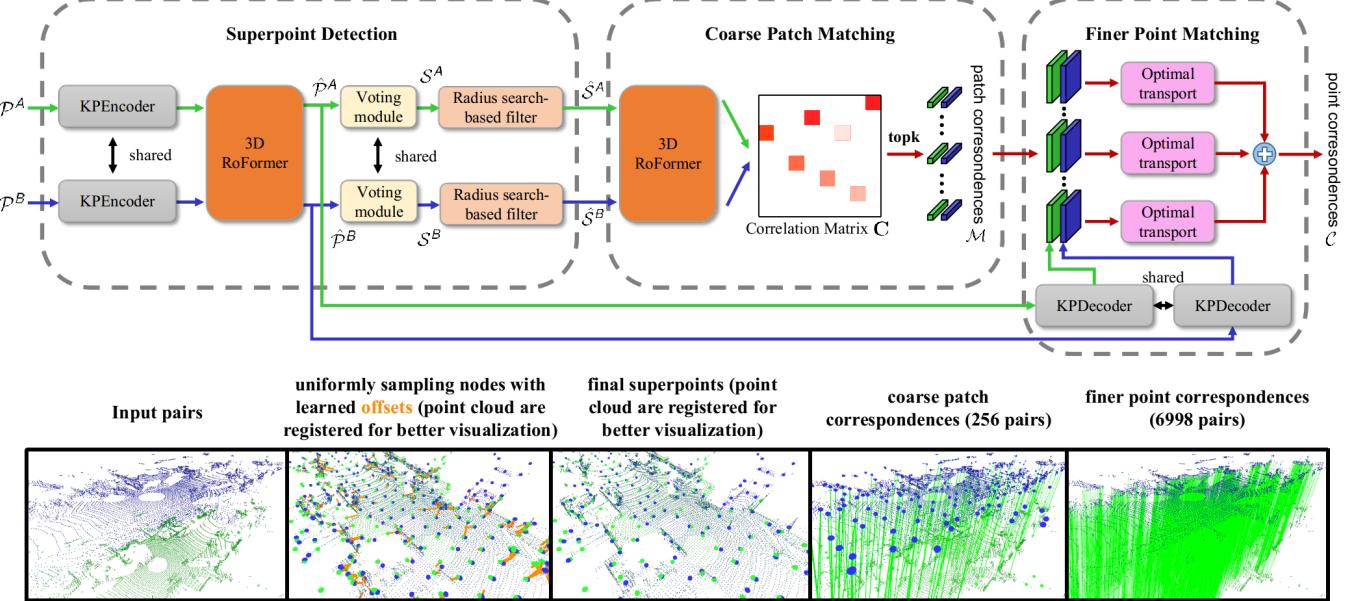


Fig. 2: **Pipeline overview.** Given two point clouds, our RDMNet first extracts superpoints from them using a superpoint detection module. Then, it applies a coarse patch matching to find the correspondences between sparse superpoints from two point clouds. Finally, a finer point matching module is then used to propagate superpoint correspondences into dense-point matches, which are used to estimate the final transformations between these two point clouds.

from the above methods, it extracts multi-level features and refines the transformation hierarchically. The quality of the sparse patch correspondences is important for such coarse-to-fine methods. To improve the correspondence accuracy, the existing methods mostly exploit the transformer network [8]. For example, CoFiNet [2] adopts the original transformer [8] as a powerful contextual information encoder to generate more accurate point correspondences. However, the vanilla transformer lacks geometric information, which hinders the performance of the position-sensitive point cloud registration. GeoTransformer [6] tackles this problem by infusing the pairwise distance and triplet-wise angular into the transformer, while NegNet [7] constructs the geometric features with point pair features [10]. Although yielding promising results, GeoTransformer results in extra-large $\mathcal{O}(n^2)$ storage complexity from the pair-wise distance and triplet-wise angular. NegNet is computationally expensive due to the normal estimation. Besides, existing works neglect the distribution of the superpoints. They use simple uniform sampling points that may separate a single object into several patches, which usually retain a low overlap ratio with patches in the other point cloud. This may lead to bad superpoint matching and dense point propagation results. Unlike the existing methods, our RDMNet uses the proposed 3D-RoFormer exploiting both the contextual and geometric information of the point cloud, which generates better correspondences for point registration.

III. OUR APPROACH

Given two point clouds $\mathcal{P}^A = \{\mathbf{p}_i^A \in \mathbb{R}^3\}_{i=1}^M$ and $\mathcal{P}^B = \{\mathbf{p}_j^B \in \mathbb{R}^3\}_{j=1}^N$, we aim to establish point correspondences between the two point clouds. To this end, we propose the RDMNet that finds correspondences in a coarse-to-fine manner. The overview of our approach is illustrated

in Fig. 2. It is built upon our devised novel 3D-RoFormer network (see Sec. III-A) and consists of three main steps: superpoint detection (see Sec. III-B), coarse patch matching (see Sec. III-C), and finer point matching (see Sec. III-D).

A. 3D-RoFormer

We first introduce our devised novel 3D-RoFormer, which is a translation-invariant transformer and the core technique of our RDMNet. We build 3D-RoFormer upon the vanilla transformer [8]. For a point \mathbf{p}_i^Q with its feature \mathbf{h}_i^Q in the query point cloud Q and all the points in the source point cloud S , the network computes the query \mathbf{q}_i , key \mathbf{k}_j , and value \mathbf{v}_j transformer feature maps with a linear projection:

$$\begin{aligned} \mathbf{q}_i &= \mathbf{W}_1 \mathbf{h}_i^Q + \mathbf{b}_1, \\ \mathbf{k}_j &= \mathbf{W}_2 \mathbf{h}_j^S + \mathbf{b}_2, \\ \mathbf{v}_j &= \mathbf{W}_3 \mathbf{h}_j^S + \mathbf{b}_3. \end{aligned} \quad (1)$$

Q, S could be downsampled input point clouds or sparse superpoints. If Q, S are the same point cloud, Eq. (1) generates the feature maps for self-attention operation, otherwise cross-attention. After that, the transformer computes an attentional weight for the query point with each source point: $\alpha_{ij} = \text{softmax}_j(\mathbf{q}_i^\top \mathbf{k}_j)$, and obtains the final attention-enhanced feature for the query point as:

$$(\tilde{\mathbf{h}}_{\text{vanilla}})_i = \sum_{j=1}^{|S|} \text{softmax}_j(\mathbf{q}_i^\top \mathbf{k}_j) \mathbf{v}_j = \sum_{j=1}^{|S|} \alpha_{ij} \mathbf{v}_j, \quad (2)$$

where $|S|$ represents the number of the points in S .

Transformer has shown to be superior for point cloud registration [2], [7], [28]. However, the vanilla transformer contains no geometric information, thus leading to suboptimal registration results. Different works are proposed to enhance

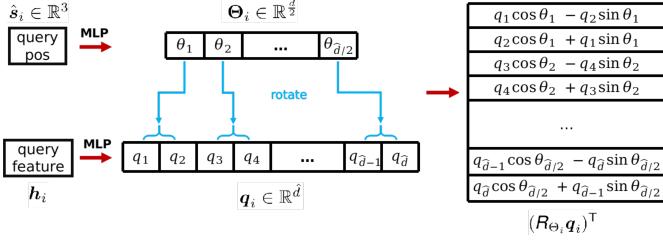


Fig. 3: The rotary 3D position embedding.

the transformer with geometric information [6], [7]. However, they are either memory-consuming [7] or time-consuming [6].

Inspired by the recent Roformer [29] using rotational information for natural language processing, we propose a novel 3D-Roformer that encodes the absolute position information with a rotation matrix for 3D point cloud registration. Based on 3D-Roformer, our RDMNet can better exploit both contextual and geometric information of the point clouds to generate more reliable keypoints. We first adapt the rotary position embedding into 3D data by leveraging a MLP and mapping the position $\hat{s}_i \in \mathbb{R}^3$ into the rotary embedding $\Theta_i = [\theta_1, \theta_2, \dots, \theta_{d/2}] \in \mathbb{R}^{\frac{d}{2}}$:

$$\Theta_i = \text{MLP}_{\text{rot}}(\hat{s}_i). \quad (3)$$

Each element in Θ_i can be treated as a rotation in a 2D plane and represented by a rotation matrix. The final formulation of the rotary 3D position embeddings $R_{\Theta_i} \in \mathbb{R}^{\tilde{d} \times \tilde{d}}$ is:

$$R_{\Theta_i} = \begin{bmatrix} \cos \theta_1 & -\sin \theta_1 & \dots & 0 & 0 \\ \sin \theta_1 & \cos \theta_1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \cos \theta_{\frac{d}{2}} & -\sin \theta_{\frac{d}{2}} \\ 0 & 0 & \dots & \sin \theta_{\frac{d}{2}} & \cos \theta_{\frac{d}{2}} \end{bmatrix}. \quad (4)$$

Applying R_{Θ_i} to a \tilde{d} dimensional vector is equivalent to divide the vector into $\tilde{d}/2$ 2D vectors and rotate each of them by $\{\theta_i | i = 1, \dots, \tilde{d}/2\}$ accordingly (see Fig. 3). We apply R_{Θ_i} and R_{Θ_j} to query q_i and key k_j respectively in self-attention operation and obtain the rotary self-attention as:

$$\alpha''_{ij} = \text{softmax}_j((R_{\Theta_i} q_i)^T R_{\Theta_j} k_j), \quad (5)$$

$$\tilde{h}_i = \sum_{j=1}^{|\hat{\mathcal{P}}|} \alpha''_{ij} v_j. \quad (6)$$

The benefits of using the proposed rotary self-attention are as follows. First, by encoding the position information as a rotation matrix, the rotary self-attention explicitly encodes the relative position information neatly. Using the properties of the rotation matrix, we can further derive Eq. (5) as:

$$\begin{aligned} \alpha''_{ij} &= \text{softmax}_j(q_i^T R_{\Theta_i}^T R_{\Theta_j} k_j), \\ &= \text{softmax}_j(q_i^T R_{\Theta_j - \Theta_i} k_j), \end{aligned} \quad (7)$$

where $\Theta_j - \Theta_i$ is naturally incorporated into the calculation of the attention scores α''_{ij} and then fused with the output feature \tilde{h}_i in Eq. (6). Therefore, our method is lightweight without requiring extra-large storage memory for relative position embeddings, which will be further verified in Sec. IV-G.

Secondly, the proposed 3D-Roformer is easy to deploy and operates very fast. Due to the sparsity of R_{Θ} , the calculation of $R_{\Theta_i} \cdot q_i$ and $R_{\Theta_j} \cdot k_j$ can be done in a computationally efficient way using vector addition and multiplication operations, for example:

$$R_{\Theta_i} \cdot q_i = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_{d-1} \\ q_d \end{bmatrix} \otimes \begin{bmatrix} \cos \theta_1 & \cos \theta_1 \\ \cos \theta_1 & \cos \theta_1 \\ \vdots & \vdots \\ \cos \theta_{\frac{d}{2}} & \cos \theta_{\frac{d}{2}} \\ \cos \theta_{\frac{d}{2}} & \cos \theta_{\frac{d}{2}} \end{bmatrix} + \begin{bmatrix} -q_2 \\ q_1 \\ \vdots \\ -q_d \\ q_{d-1} \end{bmatrix} \otimes \begin{bmatrix} \sin \theta_1 \\ \sin \theta_1 \\ \vdots \\ \sin \theta_{\frac{d}{2}} \\ \sin \theta_{\frac{d}{2}} \end{bmatrix}. \quad (8)$$

Thirdly, our proposed 3D-RoFormer is translation-invariant inherited from the linearity of MLP:

$$\Theta_j - \Theta_i = \text{MLP}_{\text{rot}}(\hat{s}_j - \hat{s}_i). \quad (9)$$

Therefore, our 3D-RoFormer will not be influenced by the changes in observation positions when used for finding correspondences. We enhance the final output features of the 3D-RoFormer $\tilde{\mathbf{H}}^A$ and $\tilde{\mathbf{H}}^B$ for point matching by interleaving the rotary self-attention and cross-attention for l times.

Benefiting from the above-mentioned advantages, our proposed RDMNet uses the devised 3D-RoFormer in both the superpoint detection and sparse patch matching modules to better find superpoint correspondences.

B. Superpoint Detection

Our approach aims first to find reliable sparse patch matches and then propagate them to dense point matches based on neighborhood consensus. Such a coarse-to-fine scheme avoids the time-consuming and unreliable global search of dense feature correspondences.

Considering a point patch as the vicinity of a keypoint, the task can be treated as the keypoint detection and vicinity grouping. We assign a keypoint for each point patch and regard them all together as one superpoint. A simple way to extract keypoints is to directly use the uniform sampling center point of each voxel [2], [7]. However, the uniformly sampled center points may separate a single object into several patches and share low overlap vicinities with center points in the other point cloud (see Fig. 1), which leads to poor patch matching and dense point match propagation in the following steps.

To address this issue, we propose the superpoint detection module to extract more reliable superpoints for the point patch partition. We use the KPEncoder [30] as the backbone of the proposed superpoint detection module that hierarchically downsamples and encodes the point cloud into the uniformly distributed nodes $\hat{\mathcal{P}}$ with associated features $\hat{\mathbf{F}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times \hat{d}}$. The node feature from KPEncoder contains only the contextual and geometric information of the single point cloud but lacks information between two point clouds, thus not being able to reason the inter-clues to make the associated superpoints compact. Therefore, we use the proposed 3D-RoFormer to fuse inter-point-cloud information and explicitly encode intra-point-cloud information at the same time. We denote the enhanced feature by 3D-RoFormer as $\tilde{\mathbf{F}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times \tilde{d}}$. A Voting module is then used to estimate the geometric offset and feature offset from the node to the proposal superpoint \mathcal{S} , i.e., $[\Delta \mathbf{P}, \Delta \mathbf{F}] = \text{Vote}(\tilde{\mathbf{F}})$, $\mathcal{S} = \hat{\mathcal{P}} + \Delta \mathbf{P}$, and $\mathbf{H} = \tilde{\mathbf{F}} + \Delta \mathbf{F}$.

We use a group of Multi-Layer Perceptron (MLP) to form the Voting module. Though very simple, it generates meaningful offsets based on the feature learned by our 3D-RoFormer (see Fig. 7) and boosts the registration performance by a large margin (see Tab. V). In this paper, we supervise the superpoints to fall in the locally significant region, which may also lead to redundant proposals located in the same significant region. Thus, we use a simple radius search-based filtering strategy to force only one proposal in the same region. We iteratively perform a radius search for each proposal and filter out the ones close to the search center. After that, we obtain the final superpoints $\hat{\mathcal{S}}$ with associated features $\hat{\mathbf{H}}$. Note that we limit the offsets $\Delta\mathbf{P}$ to a certain range, which maintains the superpoints to be evenly distributed throughout the point cloud instead of only concentrated in so-called significant areas. It also avoids possible degeneracy [5].

For each superpoint \hat{s}_i , we construct a local patch \mathcal{G}_i using a point-to-node strategy [5]. Specifically, each point is assigned to its nearest superpoint by:

$$\mathcal{G}_i = \{\mathbf{p} \in \mathcal{P} | i = \operatorname{argmin}_j (\|\mathbf{p} - \hat{s}_j\|_2), \hat{s}_j \in \hat{\mathcal{S}}\}. \quad (10)$$

There are two advantages of this strategy. First, it assigns every point to a specific superpoint without duplication or loss. Second, it adapts to different densities, which is particularly suitable for our case since our superpoints break the uniformity of the original sampling after adding offsets.

C. Sparse Patch Matching

Based on the detected superpoints, we then conduct patch matching on the coarse level and find superpoints/patch correspondences between point clouds A and B .

Since the superpoints have just been shifted and filtered by the superpoint detection module, the associated feature $\hat{\mathbf{H}}$ could be inconsistent with the surroundings. Therefore, we first feed the superpoints with the associated features to another 3D-RoFormer to update the features with the newest contextual and geometric information from the updated superpoints. Then we conduct the superpoint matching. We follow Qin *et al.* [7] and compute a Gaussian correlation matrix $\mathbf{C} \in \mathbb{R}^{|\hat{\mathcal{S}}^A| \times |\hat{\mathcal{S}}^B|}$ between normalized $\hat{\mathbf{H}}^A$ and $\hat{\mathbf{H}}^B$ with $c_{i,j} = \exp(-\|\hat{\mathbf{h}}_i^A - \hat{\mathbf{h}}_j^B\|^2)$. A dual-normalization is then performed to suppress ambiguous matches:

$$\hat{c}_{i,j} = \frac{c_{i,j}}{\sum_{k=1}^{|\hat{\mathcal{S}}^A|} c_{i,k}} \cdot \frac{c_{i,j}}{\sum_{k=1}^{|\hat{\mathcal{S}}^B|} c_{k,j}}. \quad (11)$$

We choose the largest N_c entries as the superpoint correspondences:

$$\mathcal{M} = \{(\hat{s}_{x_i}^A, \hat{s}_{y_i}^B) | (x_i, y_i) \in \operatorname{topk}_{x,y}(\hat{c}_{x,y})\}. \quad (12)$$

Based on the fast superpoint matching, we determine the corresponding matched patches and use them as the basis for the subsequent fine-level dense-point matching.

D. Dense Point Matching

On the Fine level, we aim to generate dense point matches from the coarse patch correspondences.

We leverage the KPDecoder [30] to recover point-level descriptors \mathbf{F} . Instead of using the updated superpoint feature, we recover the point-level feature from the raw anchor point feature $\tilde{\mathbf{F}}$ since there might be information loss after offsetting and filtering. For each superpoint correspondence $(\hat{s}_{x_i}^A, \hat{s}_{y_i}^B)$, we have its corresponding patch match $(\mathcal{G}_{x_i}^A, \mathcal{G}_{y_i}^B)$ and then compute a match score matrix $\mathbf{O}_i \in \mathbb{R}^{M_i \times N_i}$:

$$\mathbf{O}_i = \mathbf{F}_{x_i}^A (\mathbf{F}_{y_i}^B)^T / \sqrt{\tilde{d}}, \quad (13)$$

where $M_i = |\mathcal{G}_{x_i}^A|$ and $N_i = |\mathcal{G}_{y_i}^B|$ represent the number of points in $\mathcal{G}_{x_i}^A$ and $\mathcal{G}_{y_i}^B$ respectively.

To handle non-matched points, we append a “dustbin” row and column for \mathbf{O}_i filled with a learnable parameter $\alpha \in \mathbb{R}$. The Sinkhorn algorithm is then used to solve the soft assignment matrix $\mathbf{Z}^i \in \mathbb{R}^{(M_i+1) \times (N_i+1)}$. Different from [2], [7] that drops the dustbin and recovers the assignment by comparing the soft assignment score with a hand-tuned threshold, we directly find max entry both row-wise and column-wise on \mathbf{Z}^i which is then recovered to assignment \mathcal{C}^i :

$$\begin{aligned} \mathcal{C}^i = & \{(\mathcal{G}_{x_i}^A(m), \mathcal{G}_{y_i}^B(n)) | (m, n) \in \operatorname{toprow}_{m,n}(\mathbf{Z}_{1:M_i, 1:(N_i+1)}^i)\} \cup \\ & \{(\mathcal{G}_{x_i}^A(m), \mathcal{G}_{y_i}^B(n)) | (m, n) \in \operatorname{topcolumn}_{m,n}(\mathbf{Z}_{1:(M_i+1), 1:N_i}^i)\}, \end{aligned} \quad (14)$$

where m and n represent the indexes of the max entry of \mathbf{Z}^i .

A point is either assigned to points in the matched patch or to the dustbin. By this, we do not need manual tuning but require a discriminative assignment matrix, which can be obtained by using our proposed loss function as detailed in Sec. III-E. Note that a point is not strictly assigned to a single point in our approach, as the strict one-to-one point correspondences do not hold in practice due to the sparsity nature of the point cloud. Instead, we trust and keep the assignment results from both sides, i.e., matches from query to source and vice versa. This results in extensively more point matches while maintaining a high inlier ratio, which benefits the transformation estimation. The final correspondences are the combination of points matches from all patches:

$$\mathcal{C} = \bigcup_{i=1}^{N_c} \mathcal{C}^i. \quad (15)$$

E. Loss function and training

The final loss is the weighted sum of these three components: $L = L_s + L_c + L_f$, where L_s is the superpoint detection loss, L_c is the coarse match loss, L_f is the fine match loss.

Superpoint detection loss. The superpoint detection loss is composed of two parts $L_s = L_{s1} + L_{s2}$. The first part L_{s1} is designed to guide our superpoints lying in the significant region, and the second part L_{s2} is designed to make the superpoints close to the real measurement points. Specifically, for the first part, we do not explicitly define the significance of a point, but use a chamfer loss to minimize the distance between matched superpoints:

$$L_{s1} = \sum_{i=1}^{|\mathcal{S}^A|} \min_{\mathbf{s}_j^B \in \mathcal{S}^B} \|\mathbf{s}_i^A - \mathbf{s}_j^B\|_2^2 + \sum_{i=1}^{|\mathcal{S}^B|} \min_{\mathbf{s}_j^A \in \mathcal{S}^A} \|\mathbf{s}_i^B - \mathbf{s}_j^A\|_2^2. \quad (16)$$

Supervised by L_{s1} , we find that the superpoints tend to move to their nearest “significant” regions to minimize the distance between superpoint pairs. For the second part, we use another chamfer loss that minimizes the distance between the superpoint to its closest point:

$$L_{s2} = \sum_{i=1}^{|S^A|} \min_{p_j^A \in \mathcal{P}^A} \|s_i^A - p_j^A\|_2^2 + \sum_{i=1}^{|S^B|} \min_{p_j^B \in \mathcal{P}^B} \|s_i^B - p_j^B\|_2^2. \quad (17)$$

Coarse match loss. We follow [7] and use overlap-aware circle loss to guide the network to extract reliable superpoint correspondence with relatively high overlap. We set the patch in A with at least one positive patch in B as the anchor patches \mathcal{G}^A . For an anchor patch \mathcal{G}_i^A , its positive patch set ε_i^+ is defined as those sharing at least 10% overlap with \mathcal{G}_i^A , and its negative patch set ε_i^- is those that do not overlap with \mathcal{G}_i^A . Then the overlap-aware circle loss on A is calculated as:

$$L_c^A = \frac{1}{|\mathcal{A}|} \sum_{\mathcal{G}_i^A \in \mathcal{A}} \log[1 + \sum_{\mathcal{G}_j^B \in \varepsilon_i^+} e^{\lambda_i^j \beta_{i,j}^+ (d_i^j - \Delta^+)}, \sum_{\mathcal{G}_k^B \in \varepsilon_i^-} e^{\beta_{i,k}^- (\Delta^- - d_i^k)}], \quad (18)$$

where $d_i^j = \|\hat{\mathbf{h}}_i^A - \hat{\mathbf{h}}_j^B\|_2$, λ_i^j refers to the overlap ratio between \mathcal{G}_i^A and \mathcal{G}_j^B , and $\beta_{i,j}^+ = \gamma(d_i^j - \Delta^+)$ and $\beta_{i,k}^- = \gamma(d_i^k - \Delta^-)$ represent the positive and negative weights. The hyper-parameters setting is followed by convention: $\Delta^+ = 0.1$ and $\Delta^- = 1.4$. The overall coarse match loss is the average of overlap-aware circle loss on A and B , i.e., $L_c = (L_c^A + L_c^B)/2$

Fine match loss. To learn a discriminative soft assignment matrix and support our dense point match module, we use a gap loss on the soft assignment matrix \mathbf{Z}^i of each patch correspondence $\{\mathcal{G}_{x_i}^A, \mathcal{G}_{y_i}^B\}$. For each matched patch pair, we generate its ground truth correspondences $\mathbf{M}^i \in \{0, 1\}^{(M_i+1) \times (N_i+1)}$ with a match threshold τ , where $M_i = |\mathcal{G}_{x_i}^A|$, $N_i = |\mathcal{G}_{x_i}^B|$. The gap loss is then calculated as:

$$L_f^i = \frac{1}{M_i} \sum_{m=1}^{M_i} \log(\sum_{n=1}^{N_i+1} [(-r_m^i + \mathbf{Z}_{m,n}^i + \eta)_+ + 1]) + \frac{1}{N_i} \sum_{n=1}^{N_i} \log(\sum_{m=1}^{M_i+1} [(-c_n^i + \mathbf{Z}_{m,n}^i + \eta)_+ + 1]), \quad (19)$$

where $(\bullet)_+ = \max(\bullet, 0)$, $r_m^i = \sum_{n=1}^{N_i+1} \mathbf{Z}_{m,n}^i \mathbf{M}_{m,n}^i$ refers to the soft assignment value for the true match of m -th point in $\mathcal{G}_{x_i}^A$, and $c_n^i = \sum_{m=1}^{M_i+1} \mathbf{Z}_{m,n}^i \mathbf{M}_{m,n}^i$ refers to the soft assignment value for the true match of n -th point in $\mathcal{G}_{x_i}^B$. The final fine match loss is the average over all the matched patch pairs: $L_f = \frac{1}{2|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} L_f^i$.

We implement and evaluate our RDMNet on 4 NVIDIA RTX 3090 GPUs. The network is trained with Adam optimizer [31]. We use 5 layers of KPEncoder (4 layers of downsampling) and 3 layers of KPDecoder, which result in coarse-level points with a resolution of 4.8 m and fine-level points with a resolution of 0.6 m. The batch size is 1, and the learning rate is 10^{-4} and decay exponentially by 0.05 every 4 epochs. We also adapt the same data augmentation as in [1].

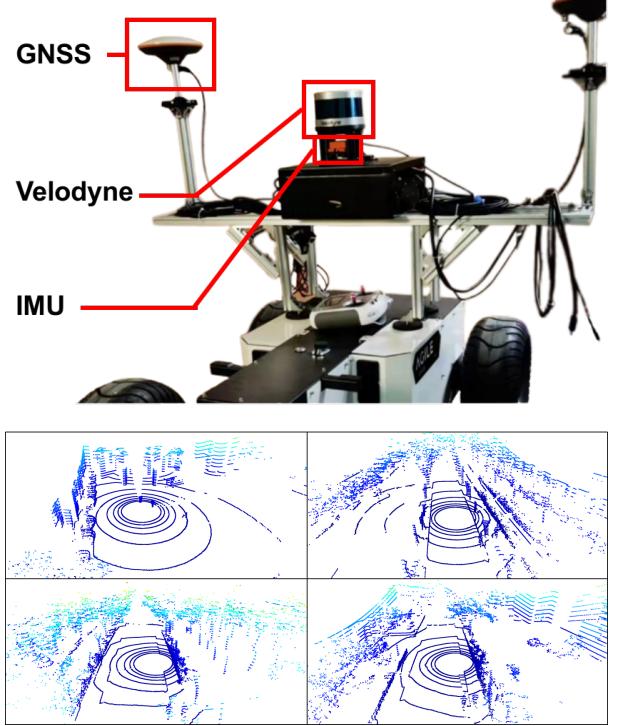


Fig. 4: Our campus data collection platform and some LiDAR data visualization of the campus dataset.

IV. EXPERIMENTAL EVALUATION

A. Dataset Overview

We evaluate RDMNet and compare it with the state-of-the-art methods on both publicly available datasets, including KITTI odometry [11], KITTI-360 [12], Apollo-SouthBay [13] and Mulran [14] datasets, and a self-recorded dataset. These datasets provide LiDAR scans collected in different environments with the corresponding ground-truth poses. The KITTI odometry and KITTI-360 contain LiDAR data collected by a Velodyne HDL64 LiDAR in Germany. These two datasets use a similar sensor setup but collect data from different times and environments. The Apollo-SouthBay dataset also uses a Velodyne HDL64 LiDAR but with a different sensor setup collecting data in the U.S. cities. The Mulran dataset contains data collected by an OS1-64 LiDAR from Korea. Fig. 4 shows our own platform equipped with a Velodyne VLP16 LiDAR, an inertial measurement unit (Xsens MTi-300), and a GNSS (INS CGI-410). We build our own dataset in a campus environment with ground-truth poses calculated by combining the GNSS and IMU with the state-of-the-art LiDAR SLAM method [32].

We follow [1], [7] and split the KITTI odometry into three sets: sequences 00-05 for training, 06-07 for validation, and 08-10 for testing. To evaluate the generalization ability, we directly apply the models trained on the KITTI odometry dataset to other datasets. Also in line with [1], [6], [7], we use the LiDAR pairs that are at most 10 m away as samples and get 1358 pairs for training, 180 pairs for validation, and 14577 pairs for testing. Note that the sensors, environments, and platform setups are different between the KITTI odometry

TABLE I: Matching results on multiple datasets under different numbers of samples. The best results are highlighted in bold, and the second bests are marked with an underline.

# Samples	KITTI			KITTI-360			Apollo			Mulran			Campus		
	5000	1000	250	5000	1000	250	5000	1000	250	5000	1000	250	2000	1000	250
<i>Feature Match Recall (%)</i>															
Predator [1]	99.64	99.64	99.64	99.87	99.78	99.44	99.98	99.98	97.86	85.96	82.01	58.76	49.12	49.12	39.18
CofiNet [2]	99.64	99.64	99.82	99.89	99.89	99.86	100	100	100	91.79	93.10	93.20	62.57	66.07	67.84
NgeNet [6]	99.64	99.64	99.64	99.94	99.94	99.94	100	100	100	<u>95.18</u>	<u>94.90</u>	87.70	91.81	88.89	70.76
Geotransformer [7]	99.82	99.82	99.82	99.89	<u>99.92</u>	99.94	100	100	100	88.28	91.05	<u>91.67</u>	98.25	98.25	97.66
RDMNet (<i>ours</i>)	99.82	99.82	99.82	<u>99.92</u>	99.94	99.94	100	100	100	98.31	98.72	98.81	100	100	100
<i>Inlier Ratio (%)</i>															
Predator [1]	62.9	50.2	29.5	60.2	48.8	29.3	44.3	31.8	16.8	14.2	11.1	6.6	5.5	5.4	4.7
CofiNet [2]	34.2	36.1	36.2	32.4	34.4	34.8	40.4	41.9	42.2	17.6	19.1	19.4	6.5	6.7	6.8
NgeNet [6]	66.5	51.5	28.6	63.2	49.6	28.5	68.6	49.2	24.8	29.1	20.7	11.1	12.4	11.4	8.1
Geotransformer [7]	<u>75.7</u>	86.0	87.5	<u>73.2</u>	83.7	85.5	<u>83.8</u>	91.0	92.4	33.6	43.7	46.3	<u>19.0</u>	<u>21.7</u>	<u>24.9</u>
RDMNet (<i>ours</i>)	86.7	93.0	95.3	84.0	91.0	93.7	92.1	96.4	97.6	<u>31.4</u>	<u>42.6</u>	51.1	34.9	36.8	41.5

TABLE II: Registration results on multiple datasets using RANSAC-50k. The results in brackets on the KITTI dataset are those reported in the original paper evaluated under bad ground truth poses. We fix it and also report the new results. The best results are highlighted in bold, and the second bests are marked with underlines. All the models are only trained on the KITTI dataset.

	KITTI	KITTI-360	Apollo	Mulran	Campus
	Registration Recall (%)				
Predator [1]	99.82	99.50	99.27	53.02	9.94
CofiNet [2]	99.82	99.62	100	80.79	36.84
NgeNet [6]	99.82	99.94	100	82.96	81.29
Geotransformer [7]	99.82	99.86	100	<u>75.68</u>	71.93
RDMNet (<i>ours</i>)	99.82	99.89	100	87.09	96.49
<i>Relative Rotation Error (°)</i>					
Predator [1]	0.25 (0.27)	0.29	0.21	1.03	1.94
CofiNet [2]	0.37 (0.41)	0.44	0.18	0.52	1.81
NgeNet [6]	0.26 (0.30)	0.30	0.18	<u>0.35</u>	1.01
Geotransformer [7]	0.22 (0.24)	0.28	0.12	0.30	0.97
RDMNet (<i>ours</i>)	0.18	0.25	0.10	0.45	0.69
<i>Relative Translation Error (cm)</i>					
Predator [1]	5.8 (6.8)	7.2	7.8	30.4	53.9
CofiNet [2]	8.2 (8.5)	10.1	6.7	17.3	38.6
NgeNet [6]	6.1 (7.4)	7.5	5.9	9.2	13.6
Geotransformer [7]	6.7 (7.4)	8.1	6.1	<u>12.0</u>	18.4
RDMNet (<i>ours</i>)	5.3	7.0	4.6	14.4	12.7

datasets to others, which thoroughly tests the generalization ability of the approaches.

B. Correspondence Matching Performance

We first evaluate the correspondence matching performance. Following [7], we use two metrics to evaluate the matching performance: i) Inlier Ratio (IR), the ratio of correct correspondences with residuals below a certain threshold, e.g., 0.6 m, after applying the ground truth transformation, and ii) Feature Match Recall (FMR), the fraction of point cloud pairs with inlier ratio above a threshold, e.g., 5%.

We compare the results of our method with the recent state-of-the-art methods: Predator [1], CofiNet [2], NgeNet [6], and Geotransformer [7]. We report the results in Tab. I with different numbers of samples. The sampling strategy is slightly different for different methods. For Predator and NgeNet, we use the default setting that samples points with probability proportional to the estimated scores. As for our approach, CofiNet, and Geotransformer, because they directly output point correspondences without interest point sampling, we pick

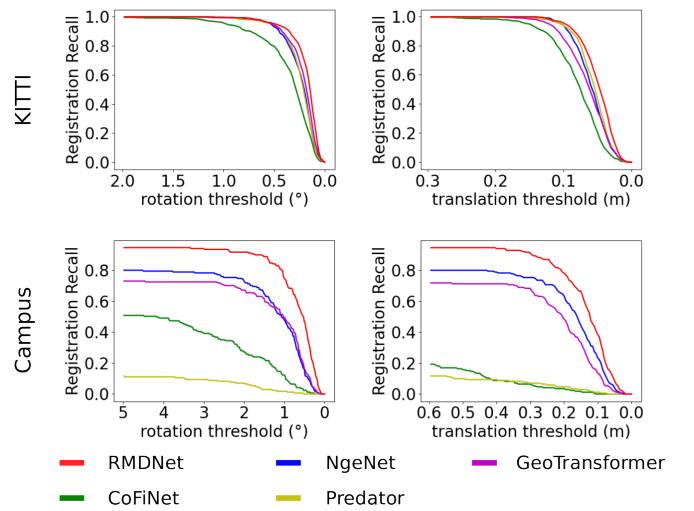


Fig. 5: The registration recall under different thresholds. When changing one of the thresholds, we fix the other one to its default number, i.e., 5° for the rotation threshold and 2 m for the translation threshold.

the top- k correspondences according to fine level soft assignment scores. Note that, the result of our approach reported on Mulran dataset is the result after removing the superpoint detection module, as we find that superpoint detection module is hard to generalize to partially occluded point clouds. For a fair comparison, we do not retrain the modified model. To our surprise, RDMNet still achieves the best FMR and the second-best IR on Mulran, as shown in Tab. I. RDMNet performs even better on other benchmarks, achieving the best FMR and IR. Notably, RDMNet exceeds the baseline by a large margin of about 10%-15% in IR. Interestingly, the performance of the keypoint-based methods and the methods that follow a coarse-to-fine manner present different trends when the number of samples becomes smaller. Keypoint-based methods, i.e., Predator and NgeNet, show a downward trend, while coarse-to-fine methods, i.e., CofiNet, Geotransformer, and ours, show an upward trend. The reason is that when the number of samples becomes smaller, the keypoint-based method becomes more difficult to sample the points in the overlap region, which is overwritten by the reliable patch correspondences for coarse-to-fine methods.

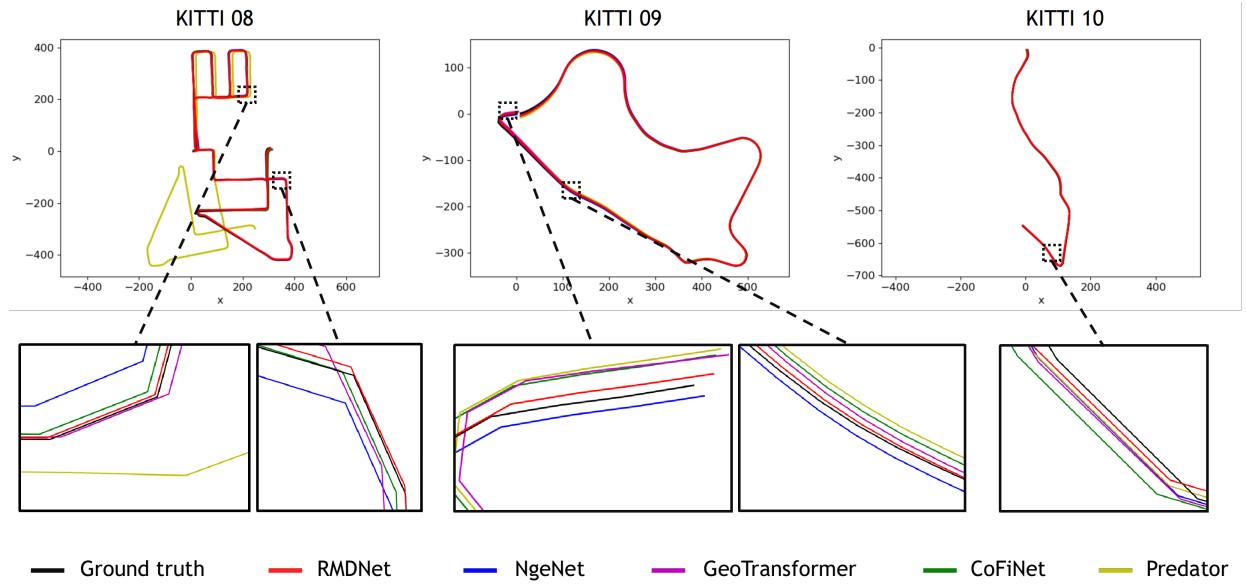


Fig. 6: Trajectory estimation results on KITTI dataset using the LiDAR pairs 10 m away.

TABLE III: Registration results on multiple datasets without RANSAC. All the models are only trained on the KITTI dataset.

	KITTI	KITTI-360	Apollo	Mulran	Campus
Registration Recall (%)					
HRegNet [27]	96.76	20.39	9.39	-	8.19
Geotransformer (LGR)	99.82	99.86	100	72.91	60.23
RDMNet (<i>ours</i> , LGR)	99.82	99.90	100	83.68	86.55
Relative Rotation Error (°)					
HRegNet [27]	1.04	2.18	2.14	-	2.72
Geotransformer (LGR)	0.31 (0.31)	0.36	0.29	0.49	0.97
RDMNet (<i>ours</i> , LGR)	0.27	0.35	0.29	0.48	0.83
Relative Translation Error (cm)					
HRegNet [27]	6.7	112.8	116.7	-	98.3
Geotransformer (LGR)	5.5 (6.0)	6.9	5.1	9.7	18.0
RDMNet (<i>ours</i> , LGR)	3.9	5.9	3.5	9.3	16.1

C. Point Cloud Registration Performance

The second experiment evaluates the point cloud registration results and supports our claim that our method outperforms the state-of-the-art method in point cloud registration.

We also follow [7] using three other metrics to evaluate the registration performance: i) Relative Translation Error (RTE), the Euclidean distance between estimated and ground truth translation vectors, ii) Relative Rotation Error (RRE), the geodesic distance between estimated and ground truth rotation matrices, and iii) Registration Recall (RR), the fraction of scan pairs whose RRE and RTE are below certain thresholds, e.g., 5° and 2 m.

We compare the results of our method with the recent RANSAC-based state-of-the-art methods: Predator [1], CofiNet [2], NgeNet [6], and Geotransformer [7] in Tab. II. There is an error in the evaluation codes of these methods using KITTI ground-truth except NgeNet. We fix the error and report the results before the fix as reported in the original papers and new results after the fix. As can be seen, Our RDMNet achieves the best RR on Mulran and outperforms all the baselines on all other datasets. Especially on the campus

TABLE IV: Absolute pose error on sequences 08-10 of KITTI dataset. Best performance is highlighted in bold while the second best is marked with an underline.

	Rot. RMSE (°)	Rot. MAE (°)	Rot. STD (°)	Trans. RMSE(cm)	Trans. MAE(cm)	Trans. STD(cm)
<i>KITTI Sequence 08</i>						
Predator [1]	104.7	39.77	8.13	10469.6	4228.3	4319.7
CofiNet [2]	8.80	4.11	1.77	879.8	335.1	381.8
NgeNet [6]	6.48	3.17	0.91	647.8	273.4	255.2
Geotransformer [7]	<u>4.18</u>	<u>2.01</u>	0.60	<u>418.3</u>	<u>171.2</u>	<u>170.3</u>
RDMNet (<i>ours</i>)	3.23	1.58	0.68	323.0	139.3	124.0
<i>KITTI Sequence 09</i>						
Predator [1]	4.14	1.99	1.12	413.5	174.9	162.5
CofiNet [2]	3.10	2.35	1.61	309.8	121.5	131.3
NgeNet [6]	4.67	2.54	0.76	466.7	173.0	206.6
Geotransformer [7]	<u>2.84</u>	1.37	0.76	<u>283.6</u>	<u>103.6</u>	<u>126.8</u>
RDMNet (<i>ours</i>)	1.91	<u>1.43</u>	0.67	190.9	76.9	79.0
<i>KITTI Sequence 10</i>						
Predator [1]	3.5	2.55	1.37	349.7	136.6	148.7
CofiNet [2]	3.2	2.9	1.6	319.9	134.6	126.5
NgeNet [6]	2.97	2.2	1.08	296.8	106.7	134.1
Geotransformer [7]	<u>1.32</u>	<u>1.1</u>	0.35	<u>132.5</u>	<u>64.5</u>	<u>41.2</u>
RDMNet (<i>ours</i>)	1.26	1.09	<u>0.39</u>	126.3	60.9	40.0

dataset, RDMNet outperforms baseline methods with a large margin on all metrics. We further evaluate RR for all the methods at different RRE and RTE thresholds on the KITTI and Campus datasets (see Fig. 5). Our RMDNet exhibits higher registration recall at all thresholds. Especially, RMDNet exceeds baseline methods by a large margin when generalizing to the campus dataset.

We also compare our method to state-of-the-art RANSAC-free methods using Local-to-Global Registration (LGR) [7]: HRegNet [27] and Geotransformer [7] in Tab. III. LGR is specifically proposed for superpoint-based approaches [7]. It calculates poses by performing weighted SVD on dense point correspondences of each patch and chooses one that admits the most inlier matches, which greatly reduces the computation time by limiting the iterations to $|\mathcal{M}|$. When using LGR, our method attains remarkable results for translation estimation

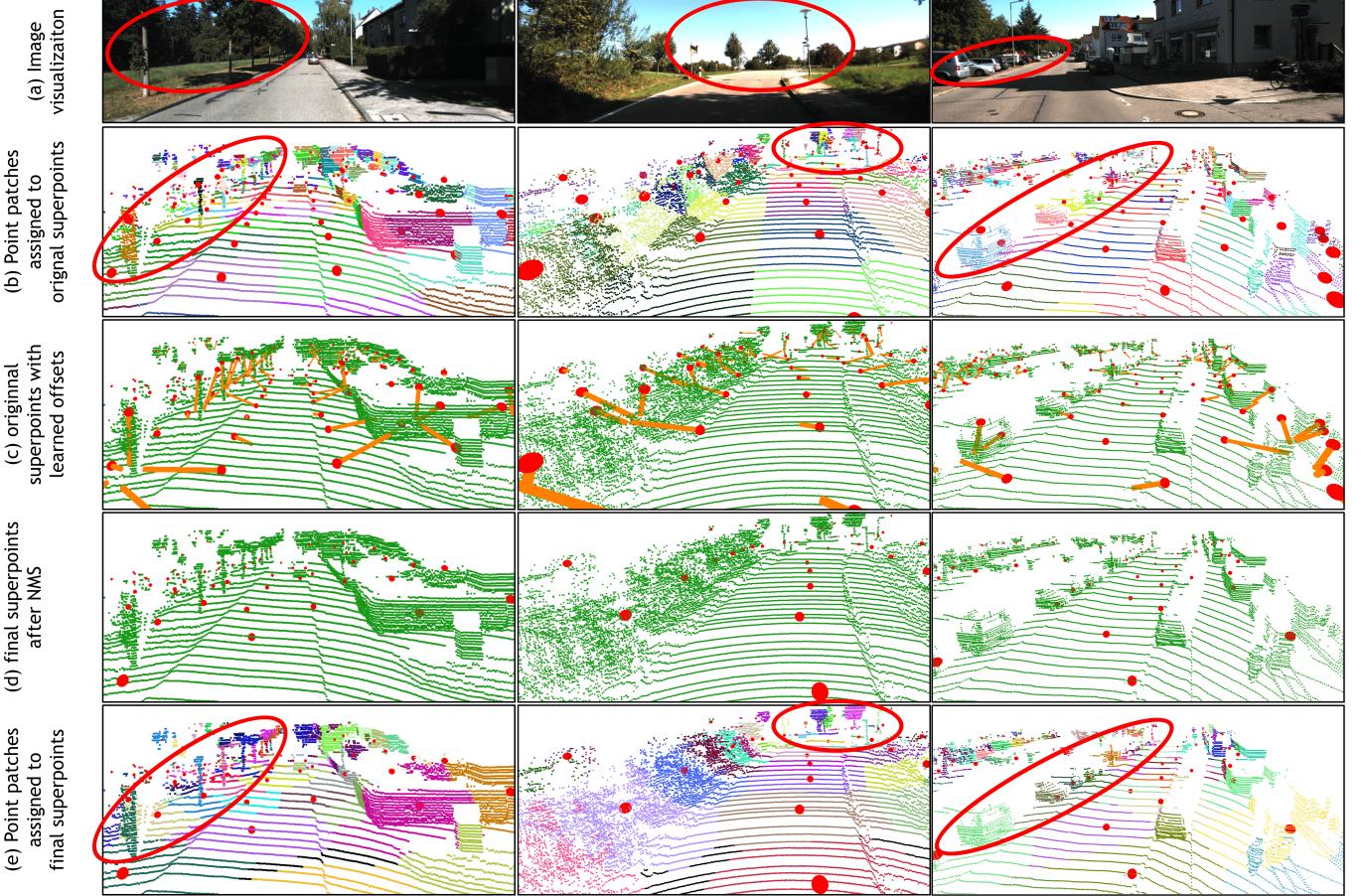


Fig. 7: Visualizations of our superpoint detection and point patch grouping. (a) shows the corresponding images of the environment only for reference. (b) shows the uniformly sampled superpoints (red dots) with their point patches grouped by point-to-node strategy. We assign each point patch a random color for visualization. (c) shows the uniformly sampled superpoints with learned offsets (orange lines). (d) shows the final superpoints after offsetting and filtering. (e) shows the point patches grouped by the final superpoints. Each point patch is assigned a random color. Uniformly sampled superpoints can easily separate a single object into several parts, as seen in (b). This poses challenges for matching. Our superpoint detection module learns a pattern that brings the superpoints close to a geometrically significant region together without using any semantic information. This leads to a more reasonable point patch grouping which benefits point matching.

and surpasses the best RANSAC-based results by about 1 cm on KITTI, KITTI-360, and Apollo datasets. In most cases, our method is the best among other RANSAC-free methods.

The above experiments evaluate the performance of all the methods in relative pose estimation. We further evaluate the absolute pose error of all the methods. We still use the LiDAR pairs 10 m away for transformation estimation and chain these transformations to obtain the trajectory. Fig. 6 shows the trajectories estimated by different methods on the three test sequences of the KITTI dataset. We calculate the root mean squared error (RMSE) and mean absolute error (MAE) of each estimated trajectory against the ground truth trajectory listed in Tab. IV. As can be seen, our RMDNet achieves overall the best performance.

So far, we have demonstrated the superior capability of our RMDNet for point cloud registration in terms of both relative and absolute pose errors. The superiorities of our method lay both in its high accuracy on the datasets similar to the training set (KITTI-360 and Apollo) and in its robustness while generalizing to the datasets under totally different sensor configurations (Mulran and Campus). The key to the robust-

ness and accuracy are the two main modules of the method, i.e., 3D-RoFormer and superpoint detection module. The 3D-RoFormer injects powerful feature representation capability into the network by encoding relative position information in a lightweight manner, while the superpoint detection module generates reliable superpoints that boost the matching and registering significantly.

D. Qualitative Results

To provide more insights into the proposed RDMNet, we visualize the superpoints and the corresponding point patches learned by our method in Fig. 7. As can be seen, compared to the vanilla uniformly distributed superpoints (second row), our method can extract the superpoints close to their nearest geometrically significant region, such as the curbs and objects, as shown from the third row to the fourth row. Based on that, our method groups the dense point patches more meaningfully, where points on a single object are gathered in the same patch (fifth row), which is important for finding accurate correspondence and obtaining good registration results.

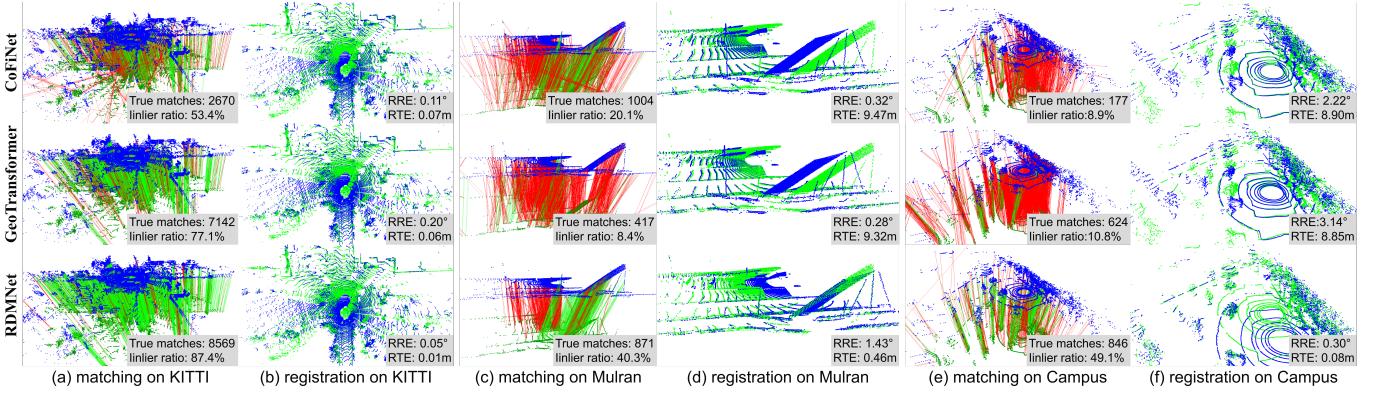


Fig. 8: Matching and registration results of our RDMNet compared to recent advances CoFiNet [2] and GeoTransformer [7]. In (a), (c), and (e), we visualize the matching on three datasets. Our RMDNet finds more inlier matches on salient regions (e.g., the curbs and the shrubs) and reject the outlier matches between similar flat patches. (b), (d), and (f) show our RMDNet achieves more accurate and robust registration compared to baseline methods.

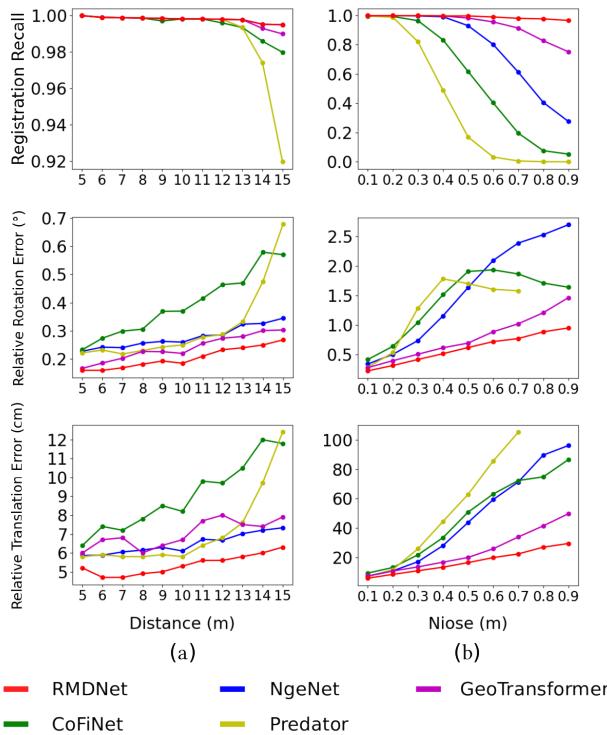


Fig. 9: Registration robustness tests in terms of (a) pair-wise distance and (b) noise level on the KITTI datasets.

Fig. 8 provides the matching and registration results of our method compared to Geotransformer [7] and CoFiNet [2]. It shows that our RDMNet finds more inlier matches on salient regions, rejects outlier matches between similar flat areas, and performs robust and accurate registration.

E. Robustness Tests

We conduct registration robustness tests regarding different overlap ratios and noise levels on the KITTI datasets. We generate the testing datasets with varying overlap ratios using the LiDAR pairs at different distances. See Fig. 9a, in terms of RR, RRE, and RTE, RDMNet achieves the best registration

TABLE V: Ablation study of individual modules.

SDM	RoPE	gap loss	KITTI			Apollo		
			RR	RRE	RTE	RR	RRE	RTE
✓	✓	✓	99.46	0.20	5.8	99.34	0.16	6.9
✓	✓	✓	99.82	0.20	5.5	98.49	0.20	6.7
✓	✓	✓	99.82	0.18	5.7	100	0.11	5.0
✓	✓	✓	99.82	0.18	5.3	100	0.10	4.6

performance for paired point clouds with varying overlap ratios. For evaluation under different noise levels, we add zero-mean Gaussian noise with σ standard deviation to the point coordinates. See Fig. 9b, RDMNet obtains the more accurate and robust registration than competitor algorithms at all noise levels. Note that RDMNet shows the superior robustness that maintains an extremely high registration recall of 96.58% compared to other baselines at a high noise level of 0.9 m.

F. Ablation Study

We conduct ablation studies on KITTI and Apollo datasets to better understand the effectiveness of each module in the proposed RDMNet, and show that the full RDMNet is the best setup. We use the model trained with negative log-likelihood loss [2], [7] using the vanilla transformer and without the superpoint detection module as the base model. Tab. V summarizes the point registration results of the ablation study. SDM refers to the superpoint detection module, and RoPE refers to the rotary position embedding. As can be seen, all modules of our method bring improvement in the point cloud registration individually. Combining all proposed modules, our RDMNet performs the best.

We also provide a study on the effectiveness of our proposed 3D-RoFormer. We compare it with other existing transformers, including vanilla transformer [2], absolute position embedding (APE) [28], and geometric embedding (GEO) [7]. We only change the transformer parts in our RDMNet while keeping the rest parts the same and comparing the point cloud registration results. As shown in Tab. VI, using our proposed 3D-RoFormer, our RDMNet achieves the best performance in all metrics on both the KITTI and Campus datasets.

TABLE VI: Ablation of 3D-RoFormer.

Model	KITTI			Campus		
	RR	RRE	RTE	RR	RRE	RTE
vanilla transformer	99.82	0.32	7.3	86.55	1.70	26.6
APE transformer	99.82	0.20	5.3	91.23	0.97	15.6
GEO transformer	99.82	0.24	6.5	88.89	1.01	20.2
3D-RoFormer (Ours)	99.82	0.18	5.3	96.49	0.69	12.7

TABLE VII: Runtime and storage of different transformer networks with different numbers of input nodes.

Model	Run time (ms)			Storage (MB)		
	1000	500	100	1000	500	100
vanilla transformer	31	24	22	40	23	6
APE transformer	33	29	24	46	24	6
GEO transformer	58	40	26	9784	2318	6
3D-RoFormer (Ours)	38	30	28	57	40	6

G. Study on Runtime and Storage

We measure the runtime and storage of the major module 3D-RoFormer for different numbers of input nodes per scan. The experiments are performed on an NVIDIA GeForce GTX 3090 GPU. As shown in Tab. VII, the runtime and storage of 3D-RoFormer increase linearly as the number of input nodes increases, similar to the vanilla transformer and APE transformer, while less than the quadratic GEO transformer.

V. CONCLUSION

In this paper, we present RDMNet that leverages a coarse-to-fine strategy to extract dense point correspondences for point cloud registration. We exploit insights from natural language processing and keypoint detection and design a novel transformation-invariant transformer named 3D-RoFormer. It learns to aggregate contextual and geometric information of the point clouds in a fast and lightweight way and extracts salient and compact superpoint pairs for point cloud registration. We evaluate and compare our approach on multiple datasets, including publicly available ones and our self-recorded dataset collected from different environments. Extensive experiments suggest that our approach outperforms the baseline methods in terms of both correspondence matching and point cloud registration with a strong generalization ability. In the future, we want to figure out why the voting module does not work well in the Mulran dataset and improve the voting module’s generalization ability. We also want to explore the potential of RDMNet to tackle the global localization problem.

REFERENCES

- [1] S. Huang, Z. Gojcic, M. Usyatsov, A. Wieser, and K. Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [2] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, 2021.
- [3] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C. Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [4] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.
- [5] J. Li and G. Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2019.
- [6] L. Zhu, H. Guan, C. Lin, and R. Han. Neighborhood-aware geometric encoding network for point cloud registration. *arXiv preprint*, 2022.
- [7] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu. Geometric transformer for fast and robust point cloud registration. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Aidan A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [9] Z. Yew and G. Lee. Rpm-net: Robust point matching using learned features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [10] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.
- [12] Y. Liao, J. Xie, and A. Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint*, 2021.
- [13] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song. L3-Net: Towards Learning Based LiDAR Localization for Autonomous Driving. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019.
- [14] X. Zhang, L. Wang, and Y. Su. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 2021.
- [15] Y. Aoki, H. Goforth, R. Srivatsan, and S. Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [16] C.R. Qi, H. Su, K. Mo, and L.J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.
- [17] Y. Zheng, Y. Li, S. Yang, and H. Lu. Global-pbnet: A novel point cloud registration for autonomous driving. *IEEE Trans. on Intelligent Transportation Systems (ITS)*, 2022.
- [18] X. Huang, G. Mei, and J. Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [19] P. J. Besl and N. D. McKay. A Method for Registration of 3D Shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.
- [20] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *Intl. Journal of Computer Vision*, 13(2):119–152, 1994.
- [21] A. Segal, D. Haehnel, and S. Thrun. Generalized-ICP. In *Proc. of Robotics: Science and Systems*, 2009.
- [22] H. Deng, T. Birdal, and S. Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] H. Deng, T. Birdal, and S. Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.
- [24] L. Li and M. Yang. Point cloud registration based on direct deep features with applications in intelligent vehicles. *IEEE Trans. on Intelligent Transportation Systems (ITS)*, 2021.
- [25] Z. Yew and G. Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.
- [26] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Proc. of the Int. Conf. on Computer Vision (ICCV) Workshops*, 2009.
- [27] F. Lu, G. Chen, Y. Liu, L. Zhang, S. Qu, S. Liu, and R. Gu. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2021.
- [28] C. Shi, X. Chen, K. Huang, J. Xiao, H. Lu, and C. Stachniss. Keypoint matching for point cloud registration using multiplex dynamic graph attention networks. *IEEE Robotics and Automation Letters (RA-L)*, 2021.
- [29] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint*, 2021.
- [30] H. Thomas, C.R. Qi, J. Deschaud, B. Marcotegui, F. Goulette, and L.J. Guibas. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2019.

- [31] D.P. Kingma and J.Ba. Adam: A method for stochastic optimization. *arXiv preprint*, abs/1412.6980, 2014.
- [32] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.