This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unlocking Generalization Power in LiDAR Point Cloud Registration

Zhenxuan Zeng[1,2] , Qiao Wu[1,2], Xiyu Zhang[1,2], Lin Yuanbo Wu[3], Pei An[4]
Jiaqi Yang[1,2*], Ji Wang[1,2*], Peng Wang[1,2]
[1]School of Computer Science, Northwestern Polytechnical University, China.
[2]Ningbo Institute, Northwestern Polytechnical University, China.
[3]Department of Computer Science, Swansea University, United Kingdom.
[4]HuaZhong University of Science and Technology, China.
{zengzhenxuan,qiaowu}@mail.nwpu.edu.cn {jqyang, j.wang, peng.wang}@nwpu.edu.cn

Abstract

In real-world environments, a LiDAR point cloud registration method with robust generalization capabilities (across varying distances and datasets) is crucial for ensuring safety in autonomous driving and other LiDAR-based applications. However, current methods fall short in achieving this level of generalization. To address these limitations, we propose UGP, a pruned framework designed to enhance generalization power for LiDAR point cloud registration. The core insight in UGP is the elimination of cross-attention mechanisms to improve generalization, allowing the network to concentrate on intra-frame feature extraction. Additionally, we introduce a progressive self-attention module to reduce ambiguity in large-scale scenes and integrate Bird's Eye View (BEV) features to incorporate semantic information about scene elements. Together, these enhancements significantly boost the network's generalization performance. We validated our approach through various generalization experiments in multiple outdoor scenes. In cross-distance generalization experiments on KITTI and nuScenes, UGP achieved state-of-the-art mean Registration Recall rates of 94.5% and 91.4%, respectively. In cross-dataset generalization from nuScenes to KITTI, UGP achieved a state-of-the-art mean Registration Recall of 90.9%. Code will be available at https://github.com/peakpang/UGP

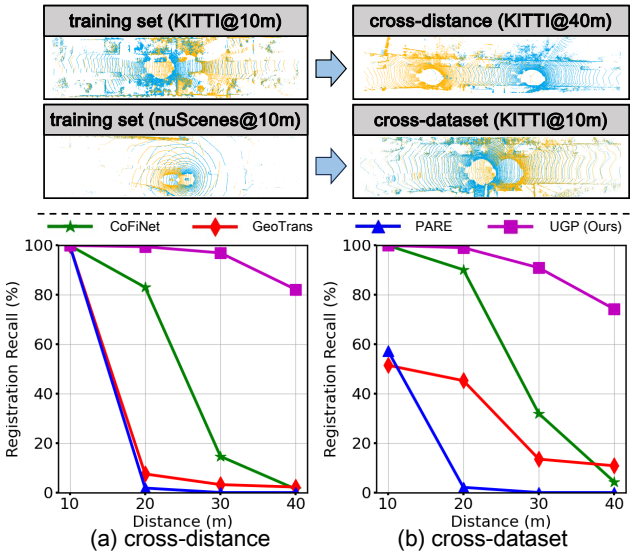


Figure 1. Illustration of the generalization performance of leading methods [28, 38, 39] in cross-distance and cross-dataset. (a) Cross-distance generalization: train on KITTI@10m, test on KITTI@40m. (b) Cross-dataset generalization: train on nuScenes@10m [4], test on KITTI@10m. These methods experience substantial performance degradation due to poor generalization. In contrast, our method achieves the best registration results on both cross-distance and cross-dataset.

1. Introduction

LiDAR point cloud registration is essential in applications such as autonomous driving [14] and SLAM [24, 25, 29]. The dynamic nature of LiDAR scenes often necessitates pairwise registration of point clouds that exhibit 1) cross-distance variations (*e.g.*, point clouds captured at different speeds or times) and 2) cross-dataset variations (*e.g.*, point clouds collected in distinct environments). To tackle these real-world challenges, this research seeks to develop a LiDAR point cloud registration framework that can effectively generalize across both cross-distance and cross-dataset variations.

Recently, learning-based methods for point cloud registration have gained prominence [1, 3, 5, 9, 13, 16, 27, 40]. However, most of these methods focus on enhancing per-

*Corresponding author

formance under same-distance or same-dataset conditions, with only a few addressing cross-dataset generalization. Notably, SpinNet [1] and BUFFER [2] use patch-wise feature extraction, which provides robustness against noise and occlusion. Nonetheless, to our knowledge, none of existing methods has systematically investigated generalization performance for LiDAR point cloud registration. Unlike evenly distributed indoor RGB-D point clouds, LiDAR point clouds exhibit uneven distribution where the point density decreases significantly as the distance from the sensor increases. This introduces unique challenges: 1) cross-distance variations alter the overlap rate between point cloud pairs and cause density changes within the overlap region, and 2) cross-dataset variations lead to differences in data characteristics (*e.g.*, 64-line vs. 32-line LiDAR) and feature distributions (*e.g.*, environmental shifts). We observe that state-of-the-art (SOTA) methods suffer noticeable performance degradation under both cross-distance and cross-dataset conditions. As shown in Fig. 1, CoFiNet [39], GeoTrans [28], and PARE [38] are coarse-to-fine methods that use the transformer to improve registration performance under the same distance or same dataset conditions. Among these approaches, cross-attention is extensively used to model the geometric consistency across the two point clouds. However, the effectiveness of this module relies on an implicit assumption: *consistency in the representation of the same structure across the two point clouds.* In dynamic scenarios where cross-distance and cross-dataset generalization challenges arise, this implicit assumption often does not hold due to the inconsistent density distribution of LiDAR point clouds.

To tackle these challenges, we propose a LiDAR point cloud registration framework with strong generalization power (**UGP**). The framework's key insight is eliminating the cross-attention module to **U**nlock **G**eneralization **P**ower and enable the network to focus on intra-frame feature extraction. First, to reduce the impact of inconsistent geometric representations in inter-frame, we eliminate the cross-attention module. This encourages the network to concentrate on intra-frame spatial information. Second, to reduce feature ambiguity in LiDAR scenes and capture finer local feature associations, we introduce a progressive self-attention module. This module gradually increases the attention range of superpoints, enabling them to prioritize local spatial information and construct a multi-scale spatial representation. Lastly, recognizing the importance of scene-element-level (*e.g.*, roads, corners) semantic information in reducing scene ambiguity, we incorporate Bird Eye View (BEV) features to improve accuracy during the coarse matching stage. In summary, our main contributions are as follows:

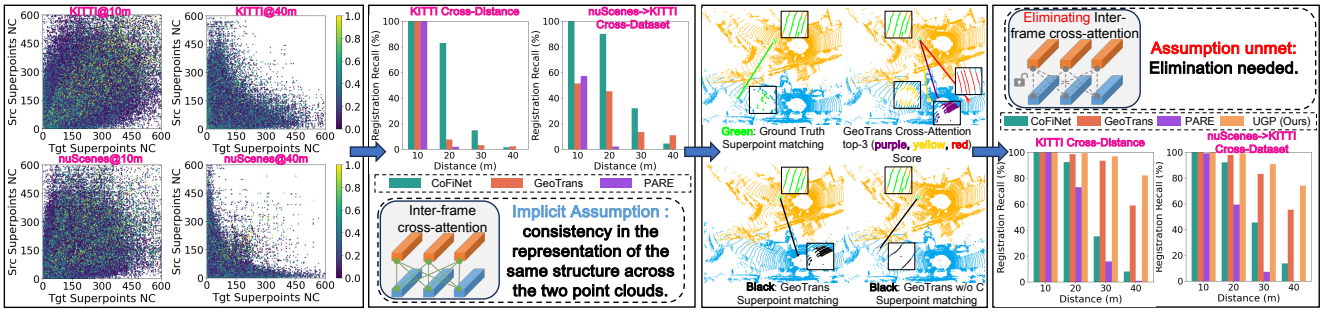
- We reveal that inconsistent geometric representations in LiDAR scenarios cause the cross-attention module to limit network generalization capability. An effective way to unlock this is to eliminate the cross-attention module.
- We also propose a progressive self-attention module and a BEV feature extraction module to more effectively focus on fine local space associations and capture semantically rich information, reducing ambiguity in point clouds and enhancing generalization.
- In cross-distance generalization from 10m to 40m, UGP achieved 82.0% (+20.8%) registration recall (RR) on KITTI and 72.3% (+57.1%) RR on nuScenes, outperforming BUFFER [2]. In cross-dataset generalization from nuScenes to KITTI, our method achieved a state-of-the-art mean RR of 90.9% (+6.2%) over BUFFER [2].

2. Related Work

Learning-based Point Cloud Registration. Learning-based point cloud registration methods have become a powerful approach for aligning point clouds. Current methods often employ a coarse-to-fine strategy [6, 23, 28, 34, 38–41], which loosens the constraints of strict point-to-point matching in favor of patch-level matching. This approach reduces repeatability requirements, significantly enhancing registration accuracy. However, these methods rely heavily on the data distribution of training point cloud pairs, which limits their generalization capability. Following this line, point-wise methods [3, 8] utilize architectures such as Minkowski [7] and KPConv [32] to extract dense point features, while patch-wise methods [1, 2, 9, 13, 27] capture features from the local regions around key points through carefully designed networks. The robustness of local patches to noise and occlusion gives these patch-based approaches a promising generalization potential. Recent studies [18–20] have focused on point cloud registration across varying distances. However, these approaches rely solely on visible full-range data and require carefully designed multi-frame training to ensure registration performance within the visible range.

Point Cloud Registration with Transformers. The Transformer architecture [33] introduced an attention mechanism that captures global dependencies and has been widely used for point cloud registration. For example, CoFiNet [39] uses vanilla attention mechanism to establish global geometric consistency, while GeoTrans [28] integrates geometric structure to enhance feature robustness. PEAL [41] employs one-way attention with overlap priors to reduce ambiguity from non-overlapping points. DCATr [6] limits its attention to specific cues to minimize irrelevant interference. Unlike these methods, we propose a progressive self-attention mechanism that gradually expands the spatial attention range, prioritizing finer local spatial structure associations, reducing ambiguity in large scenes, and improving registration accuracy.

Multi-modal Point Cloud Registration. Incorporating



(a) LiDAR Point Cloud Registration Characteristics (b) Existing SOTA methods show limitations (c) Preliminary validation (d) Eliminated validation

Figure 2. **Motivation.** We conducted a statistical analysis of the matching characteristics in LiDAR scenes and identified the limitations of existing methods. This was followed by preliminary visualizations and elimination experiments to validate our findings. (a) Each point in the figure represents a ground truth corresponding superpoint pair. The position of each point indicates the neighborhood count (NC) of the superpoint within a radius of $r = 2.4m$ in both the source (src) and target (tgt) point clouds, and the color represents the overlap degree of the corresponding superpoint pairs after rotation by the ground truth transformation. (b) Performance of existing methods in cross-distance (upper left) and cross-dataset (upper right). (c) GeoTrans [28] cannot match superpoints (cross-distance, from KITTI@10m to KITTI@20m). (d) The performance of existing methods after eliminating cross-attention, and UGP (ours).

multi-modal information for point cloud understanding has gained significant traction in recent research. For example, RPVNet [35] enhances segmentation accuracy by integrating the range image modality. Methods such as BYOC [10], PCRCG [43], and IGReg [36] fuse image features with point clouds to support registration tasks. Similarly, PEAL2D [41] utilizes a 2D image to derive an overlap prior, aiding in point cloud registration. BEV images, widely adopted in 3D object detection [37] and place recognition [22], are particularly effective due to their clear global edge features. To enhance matching accuracy in the coarse stage, we explicitly incorporate BEV images into point cloud registration.

3. Motivation

In real-world dynamic scenarios, we face the challenge of registering point clouds with 1) cross-distance variations (*e.g.*, point clouds captured at different speeds or times) and 2) cross-dataset variations (*e.g.*, point clouds collected in distinct environments or using different LiDAR types). This challenge requires the LiDAR point cloud registration framework to possess cross-distance and cross-dataset generalization power. To address this, we analyze the matching characteristics under both challenges and find that existing methods have limited performance. Finally, we propose targeted improvements that effectively address these existing challenges.

a) Inherent Challenges with LiDAR Point Cloud Registration. Fig. 2 (a) highlights the challenges in LiDAR point cloud registration by visualizing data distributions of ground truth matching point pairs across varying distances and datasets. 1) At closer distances (*e.g.*, KITTI@10m or nuScenes@10m), there is a high number of matched pairs with consistent density (clustered near $y = x$) and a high overlap score. As the distance increases (*e.g.*, KITTI@40m vs. KITTI@10m or nuScenes@40m vs. nuScenes@10m), the proportion of pairs with inconsistent density grows. Additionally, increased distance brings changes in perspective and occlusion, reducing overlap rates even among points with similar density. 2) Cross-dataset comparisons reveal differences in data characteristics, such as the shift from nuScenes (32-line LiDAR) to KITTI (64-line LiDAR).

b) Implicit Assumption of Cross-Attention Based Methods. The effectiveness of SOTA methods such as CoFiNet [39], GeoTrans [28], and PARE [38] relies on an implicit assumption: *consistency in the representation of the same structure across two point clouds*. However, cross-attention cannot adapt to variations in the consistency representation of the same structure across different distances and datasets, limiting the network's generalization performance, as shown in Fig. 2 (b).

c) Cross-Attention Elimination. To demonstrate the limitations of cross-attention, we present results from GeoTrans [28] trained on KITTI@10m and tested on KITTI@20m, as shown in Fig. 2 (c). First, we show a challenging superpoint matching example on the ground, where uneven LiDAR distribution and limited geometric detail make accurate matching difficult. As the distance between point cloud pairs increases (*e.g.*, green ground-truth matches at 20m), variations in density and occlusion worsen matching accuracy. Cross-attention-based methods like GeoTrans [28], fail in cross-distance matching by relying on density-similar but incorrect matches. Visualization of GeoTrans [28]'s top-3 score areas from its final cross-attention layer shows that it primarily captures similarly

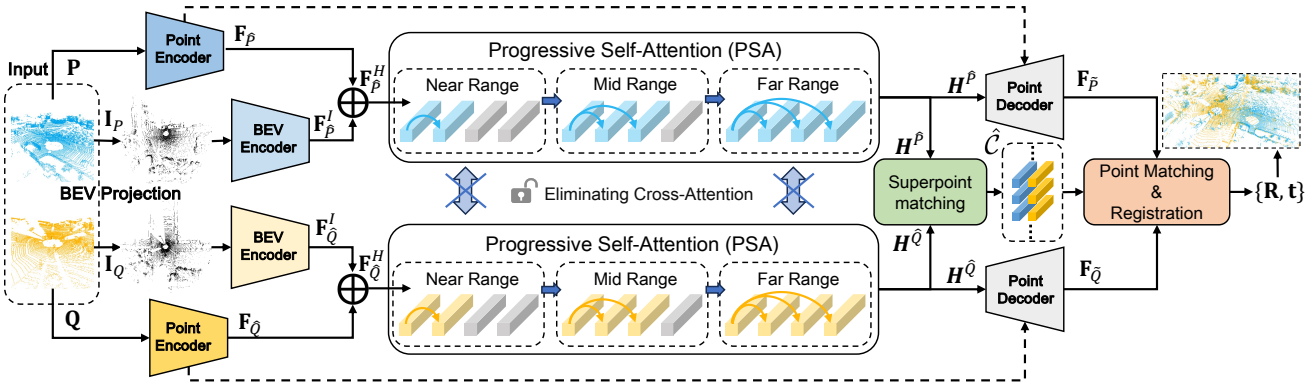


Figure 3. **Overview of UGP.** The input point cloud is first projected to obtain the corresponding BEV image. Then, the point cloud and BEV image are fed into the Point Encoder and BEV encoder, respectively, for downsampling and feature extraction. In the superpoint matching stage, the indexing relationship between superpoints and BEV is used to fuse point-level features and image features. The fused features are used as input to the progressive self-attention module to extract robust and consistent intra-frame features. Finally, the superpoint matching results are propagated to dense points for dense matching, enabling the recovery of the rigid transformation.

dense structures and struggles to adapt to shifts in data distribution, resulting in outlier matches. By eliminating cross-attention, superpoint matching refocuses on correct areas near the ground truth. Preliminary validation followed by elimination experiments (Fig. 2 (d)) confirms that eliminating cross-attention greatly enhances model generalization.

d) Boosting the Robustness and Consistency of Intra-Frame Feature Learning. After eliminating the cross-attention module, we encourage the network to concentrate on intra-frame spatial information. A vanilla self-attention [33] forces each point to interact with all other points equally. However, this leads to feature ambiguity due to irrelevant distant points. To improve the robustness and consistency of intra-frame feature extraction, we propose a progressive self-attention. In the initial layer, the attention range is restricted to the local space, preserving fine-grained local information. In subsequent layers, the attention range is gradually expanded to incorporate the global context. This progressive approach reduces feature ambiguity and enhances the model's ability to effectively capture both local and global structure. Besides, point-based backbones such as KPConv [32] are difficult to establish a association between local geometric structures and global background information [21, 35]. This further worsens the ambiguity of the features. To this end, we introduce BEV which can provide additional semantic information of scene elements (*e.g.*, roads, corners) to the local geometric features extracted by KPConv.

4. Method

Given two sets of partially overlapped point clouds $\mathbf{P} \in \mathbb{R}^{N\times 3}$ and $\mathbf{Q} \in \mathbb{R}^{M\times 3}$, our objective is to recover the optimal rigid transformation between them, denoted as $\mathbf{T} = \{\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3\}$, where $\mathbf{R}$ represents rotation and $\mathbf{t}$ represents translation. As illustrated in Fig. 3, we adopt a coarse-to-fine strategy and eliminate the cross-attention module. First, the point cloud is directly projected to obtain the BEV image (Sec. 4.1). Then, the Point-Encoder and BEV-Encoder are used to extract the scale features of the point cloud and BEV image. Next, the patch features of the point cloud and image are fused to obtain the superpoint features (Sec. 4.2). The superpoint matching module eliminates the cross-attention module and utilizes only the progressive self-attention to extract superpoint features, thereby obtaining superpoint correspondences (Sec. 4.3). Finally, we follow GeoTransformer [28], using the same point matching module to refine the superpoint matching and recover the alignment transformations with the LGR [28] approach (Appendix Sec. 7).

4.1. BEV Projection

In LiDAR point cloud registration, the significant global edge and texture feature are important for superpoint matching. From a 2D perspective, the BEV representation of the point cloud provides a global view of the environment and offers important semantic information about scene elements. These prominent features reduce the ambiguity of the scene and improve feature consistency.

Given a point cloud $\mathbf{P} = \{(x_i, y_i, z_i) \mid i = 1, 2, \ldots, N\}$, where (x_i, y_i, z_i) represents the coordinates of each point, we project into a BEV image $\mathbf{I} \in \mathbb{R}^{H\times W}$, where H and W denote the height and width of the BEV image, respectively. Assuming that the ground area covered by this aerial view is $[x_{\min}, x_{\max}]$ and $[y_{\min}, y_{\max}]$, the projection process can be expressed by the following formula :

$$u_i = \left\lfloor \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \cdot H \right\rfloor, v_i = \left\lfloor \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \cdot W \right\rfloor, \quad (1)$$

where (u_i, v_i) is the pixel coordinate of point (x_i, y_i, z_i) in

the BEV image. To capture the global structure, we fill each pixel position with a gray value of $\mathbf{I}(u_i, v_i) = 1$.

4.2. Superpoint Feature Extraction and Fusion

Our feature extraction module consists of a Point-Encoder and a BEV-Encoder. The Point-Encoder employs a multi-level KPConv structure [32], with voxel downsampling applied at each layer to decrease the number of points and ultimately extract superpoint features. The BEV-Encoder uses a multi-level ResNet structure [15], applying 2D max pooling at each layer to downsample the image and extract patch-level features. Finally, we fuse the two features to obtain richer superpoints feature expressions.

Specifically, given the input point clouds $\mathbf{P}$ and $\mathbf{Q}$, along with their corresponding BEV images $\mathbf{I}_P$ and $\mathbf{I}_Q$, the Point-Encoder outputs the downsampled superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, as well as their corresponding features $\mathbf{F}_{\hat{P}} \in \mathbb{R}^{\hat{n} \times \hat{d}}$ and $\mathbf{F}_{\hat{Q}} \in \mathbb{R}^{\hat{m} \times \hat{d}}$, where $\hat{d}$ denotes the feature dimension of each superpoint. At the same time, the BEV-Encoder generates the patch features $\mathbf{F}^I_{\hat{P}} \in \mathbb{R}^{H' \times W' \times d'}$ and $\mathbf{F}^I_{\hat{Q}} \in \mathbb{R}^{H' \times W' \times d'}$, where H' and W' represent the height and width of the downsampled BEV image, and d' denotes the feature dimension for each patch in the BEV image. Afterwards, feature fusion is performed based on the indexing relationship between superpoints and BEV patches, where the features are concatenated to combine 3D geometry and 2D texture information. This results in the fused features $\mathbf{F}^H_{\hat{P}}$ and $\mathbf{F}^H_{\hat{Q}}$ for the superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, respectively.

4.3. Superpoint Matching Module

For the coarse-to-fine method, superpoint matching accuracy directly determines final registration performance.

Eliminating Cross-Attention to Concentrate on Intra-Frame Feature Learning. The distributional characteristics of LiDAR data differ from indoor RGB-D point clouds, causing inconsistencies in geometric representations across distances and datasets and limiting network generalization. To improve the framework's generalization performance, we adopt a simple, effective approach based on the observations in Sec. 3: Eliminating the cross-attention module and retaining self-attention to focus on intra-frame feature extraction, enhancing generalization.

In the following, we describe the process for computing the feature for superpoint $\hat{\mathbf{P}}$ with self-attention, which similarly applies to $\hat{\mathbf{Q}}$. Given the input feature matrix $\mathbf{X} \in \mathbb{R}^{\hat{n} \times d_t}$, where d_t represents the dimension of each input feature, the output feature matrix $\mathbf{Z} \in \mathbb{R}^{\hat{n} \times d_t}$ is obtained as the weighted sum of the projected input features:

$$\mathbf{z}_i = \sum_{j=1}^{\hat{n}} a_{i,j} \left(\mathbf{x}_j \mathbf{W}^V \right), \tag{2}$$

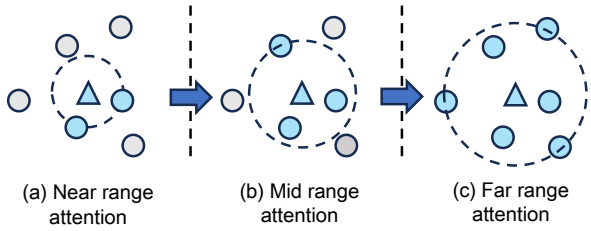


Figure 4. Illustration of our proposed progressive self-attention. In the initial layer, the triangular points prioritize attention to the local space around them, and the attention range is gradually expanded in subsequent layers

where the attention weights $a_{i,j}$ are obtained by applying a row-wise softmax to the attention scores $e_{i,j}$, and the attention scores $e_{i,j}$ computed as:

$$e_{i,j} = \frac{\left(\mathbf{x}_i \mathbf{W}^Q \right) \left(\mathbf{x}_j \mathbf{W}^K + \mathbf{r}_{i,j} \mathbf{W}^R \right)^T}{\sqrt{d_t}}. \tag{3}$$

Here, $\mathbf{r}_{i,j}$ refers to the geometric structure embedding proposed in GeoTransformer [28]. The terms $\mathbf{W}^Q$,$\mathbf{W}^K$,$\mathbf{W}^V$,$\mathbf{W}^R \in \mathbb{R}^{d_t \times d_t}$ are the respective projection matrices for queries, keys, values and geometric structure embeddings.

Progressive Self-Attention Module. In real-world scenarios, local contextual information plays a crucial role in point cloud feature extraction. Existing methods [6, 28, 38, 39, 41] apply global self-attention directly to the input point cloud features. Although self-attention can capture long-range global dependencies, this approach is more susceptible to noise in large-scale scenes, which aggravates the ambiguity in feature matching. To reduce the matching ambiguity in LiDAR scenes, we introduce a progressive self-attention module, as shown in Fig. 4. This module dynamically adjusts the attention range in a progressive manner, based on the distances between each point and other points in the global context. In the initial layer ($L = 1$), the module targets local association refinement by focusing on nearby superpoints, effectively reducing the ambiguity that can arise from initial global attention. In subsequent layers ($L > 1$), the attention range is gradually expanded, enabling the model to build multi-scale spatial representations of superpoints. This stepwise approach enhances the robustness of superpoint features.

Next, we will detail the computation process for $\hat{\mathbf{P}}$ at layer $L = k$, which similarly applies to $\hat{\mathbf{Q}}$. To implement progressive, the masked attention scores $e^{(L=k)}_{i,j}$ are computed as:

$$e^{(L=k)}_{i,j} = \mathbf{M}^{(L=k)}_{i,j} \cdot e_{i,j}, \tag{4}$$

where $\mathbf{M}^{(L=k)}_{i,j}$ is a layer-specific mask designed to filter attention scores based on the Euclidean distance between

Model	Train on KITTI@10m					KITTI@20m					KITTI@30m					KITTI@40m					
	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	mRR
FCGF [8]	**0.213**	0.064	0.864	0.185	98.9	**0.358**	**0.121**	5.812	1.477	92.1	1.051	0.339	11.706	10.514	64.1	1.477	0.767	19.855	30.598	18.7	68.5
SpinNet [1]	0.515	0.104	0.948	0.322	97.7	1.436	0.237	9.418	7.833	58.0	2.018	0.638	22.107	28.821	1.1	–	–	23.191	39.544	0.0	39.2
Predator [16]	0.282	0.070	0.282	0.076	**99.8**	1.602	0.324	16.438	12.403	30.2	1.903	0.419	28.594	28.054	1.1	–	–	30.605	38.637	0.0	32.8
CoFiNet [39]	0.344	0.083	0.357	0.089	**99.8**	1.133	0.256	3.758	2.924	82.9	1.862	0.557	18.521	24.992	14.6	2.054	0.807	25.716	39.071	1.4	49.7
GeoTrans [28]	0.244	0.069	0.301	0.075	**99.8**	1.640	0.395	26.555	24.533	7.5	1.355	0.503	45.092	34.987	3.2	1.431	0.618	46.972	45.065	2.2	28.2
BUFFER [2]	0.259	0.072	0.262	0.079	**99.8**	0.437	0.131	0.475	0.345	98.6	0.719	**0.230**	2.290	1.910	93.5	1.050	**0.364**	14.221	15.811	61.2	88.3
PARE [38]	0.239	**0.054**	**0.242**	**0.061**	**99.8**	1.284	0.401	19.192	27.340	1.8	–	–	23.682	37.837	0.0	–	–	29.139	47.570	0.0	25.4
CoFiNet w/o C	0.366	0.085	0.379	0.109	**99.8**	0.813	0.182	2.305	1.504	92.5	1.734	0.490	14.172	18.061	35.1	1.970	0.785	27.077	36.130	7.9	58.8
GeoTrans w/o C	0.238	0.066	0.295	0.073	**99.8**	0.426	0.146	0.501	0.225	98.6	0.699	0.285	4.808	1.889	93.5	**0.949**	0.476	32.318	14.096	59.0	87.7
PARE w/o C	0.289	0.071	0.292	0.078	**99.8**	0.924	0.288	9.548	4.627	73.0	1.330	0.574	38.783	23.824	15.7	3.660	1.456	47.526	39.225	0.7	47.3
UGP(Ours)	0.245	0.071	0.298	0.078	**99.8**	0.393	0.147	**0.449**	**0.221**	**99.3**	**0.677**	0.290	**2.127**	**0.805**	**96.8**	0.975	0.494	**13.095**	**6.168**	**82.0**	**94.5**
Model	**Train on nuScenes@10m**					**nuScenes@20m**					**nuScenes@30m**					**nuScenes@40m**					
	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	mRR
FCGF [8]	0.338	**0.176**	0.396	0.374	97.4	0.511	0.343	2.524	3.105	84.4	1.015	0.492	6.801	12.178	53.5	1.560	0.519	12.496	28.355	19.7	63.8
Predator [16]	0.365	0.178	0.365	**0.178**	**100**	1.640	**0.312**	22.617	10.038	15.8	1.855	**0.279**	25.102	20.76	5.3	1.809	0.175	31.424	29.966	5.1	31.6
CoFiNet [39]	0.461	0.195	0.625	0.416	97.0	1.370	0.482	18.662	7.398	58.0	1.810	0.556	33.388	23.109	12.7	1.923	0.365	32.966	33.459	6.5	43.6
GeoTrans [28]	**0.259**	0.178	**0.259**	**0.178**	**100**	0.733	0.412	9.099	5.059	57.1	0.734	0.382	22.755	19.452	9.6	0.571	0.294	29.104	29.317	7.2	43.5
BUFFER [2]	0.340	0.178	0.390	0.394	97.5	0.528	0.347	0.866	2.917	84.7	0.918	0.523	9.901	14.183	46.6	1.073	0.567	25.865	30.811	15.2	61.0
PARE [38]	0.449	0.190	0.475	0.202	99.7	0.906	0.445	6.834	9.638	9.2	**0.674**	0.365	10.530	19.084	5.3	**0.346**	**0.163**	13.746	28.349	4.6	29.7
UGP(Ours)	0.270	0.179	0.270	0.179	**100**	**0.467**	0.338	**0.646**	**0.415**	**99.4**	0.780	0.460	**1.284**	**1.456**	**93.9**	1.142	0.636	**10.532**	**8.641**	**72.3**	**91.4**

Table 1. **Cross-distance generalization experiments.** For both the KITTI and nuScenes datasets, **we train at 10m and then test at 10m and farther distances at 20m, 30m, and 40m.** RRE and RTE denote the error for successfully matched point cloud pairs, while RRE* and RTE* reflect the error for all point cloud pairs, providing a more comprehensive evaluation. The final column shows the mean Registration Recall.

points. The mask $\mathbf{M}_{i,j}^{(L=k)}$ is computed by dividing the maximum distance between superpoint $\hat{p}_i$ and all other superpoints $\hat{p}_j$ in the $\hat{\mathbf{P}}$ into S segments. For a given layer $L = k$, the mask is defined as:

$$\mathbf{M}_{i,j}^{(L=k)} = \begin{cases} 1, & \text{if } d_{i,j} \leq \frac{k \cdot d_{i,\max}}{S},\ k = 1, 2, \ldots, S \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $d_{i,\max} = \max_{i,j \in \hat{\mathbf{P}}} d_{i,j}$, k represents the segment index, with $k = 1$ corresponding to the smallest distance range and $k = S$ corresponding to the largest. This enables the model to gradually expand the receptive field, focus on more refined local space in the early stage, reduce distant noise, and then learn multi-scale spatial feature expressions. After performing S iterations of attention, the mixed superpoint features $\mathbf{H}^{\hat{P}}$ and $\mathbf{H}^{\hat{Q}}$ are obtained.

Superpoint Matching. To obtain the superpoint correspondences, we follow [28]. First, the features $\mathbf{H}^{\hat{P}}$ and $\mathbf{H}^{\hat{Q}}$ obtained in the coarse stage are normalized to the unit sphere to yield the features $\bar{\mathbf{H}}^{\hat{P}}$ and $\bar{\mathbf{H}}^{\hat{Q}}$. Next, we compute a Gaussian correlation matrix $\hat{\mathbf{S}}$ with $\hat{s}_{i,j} = \exp\left(-\left\|\bar{\mathbf{h}}_i^{\hat{P}} - \bar{\mathbf{h}}_j^{\hat{Q}}\right\|_2^2\right)$. Following this, we perform dual normalization on $\hat{\mathbf{S}}$ to obtain $\bar{\mathbf{S}}$. Finally, based on the normalized results, the top-k point pairs with the highest correlation scores are selected as superpoint correspondences $\hat{\mathcal{C}} = \left\{(\hat{p}_i, \hat{q}_j) \mid (\hat{p}_i, \hat{q}_j) \in \text{topk}\left(\bar{\mathbf{S}}\right)\right\}$.

5. Experiments

5.1. Experimental Settings

Datasets. We evaluated our method on two outdoor LiDAR datasets: KITTI (64-line LiDAR) [12] and nuScenes (32-line LiDAR) [4]. Following previous works [8, 16, 39], sequences 0-5 are used for training, 6-7 for validation, and 8-10 for testing, with pairs created from scans at least 10m apart. To evaluate cross-distance and cross-dataset generalization, we expanded the test set to include pairs separated by 20m, 30m, and 40m. The goal is to train on the KITTI@10m dataset and test generalization on KITTI@20m, KITTI@30m and KITTI@40m. For nuScenes, we follow the common protocol [4], splitting it into training, validation, and testing subsets. Using the same approach, we train and validate on nuScenes@10m and test on nuScenes@20m, nuScenes@30m, and nuScenes@40m. To verify cross-dataset generalization, we trained on nuScenes@10m and tested on KITTI@10m, KITTI@20m, KITTI@30m, and KITTI@40m. Additional experiments are in Appendix Sec. 9.

Metrics. We follow the previous work [6, 16, 28] and use five evaluation metrics: Patch Inlier Ratio (PIR), Inlier Ratio (IR), Relative Rotation Error (RRE), Relative Translation Error (RTE), Registration Recall (RR), and mean Registration Recall (mRR). RR in all experiments is based on the criteria RRE $< 5°$ and RTE $<$ 2m. The definitions of the metrics can be found in Appendix Sec. 8.

5.2. Generalization Experiment

Cross-Distance Generalization. We compare the proposed UGP with other state-of-the-art methods in KITTI and

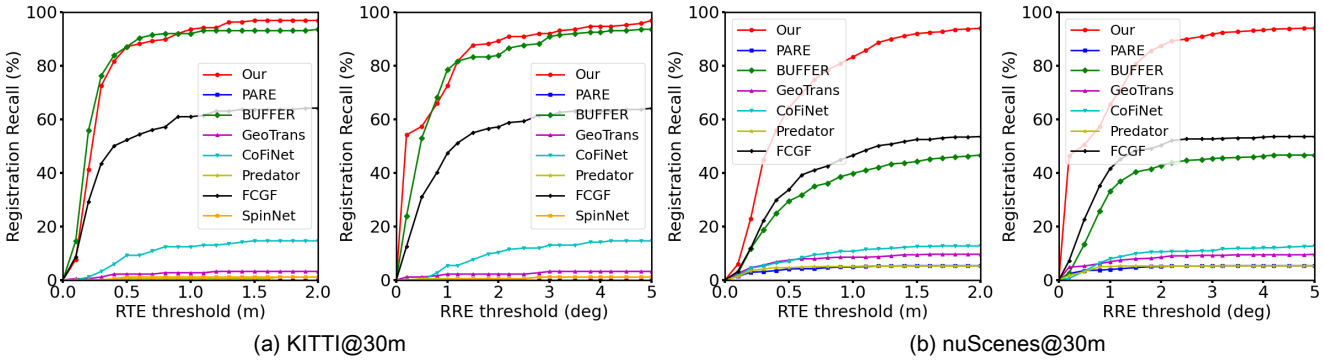


Figure 5. Cross-distance registration recall results of different RRE and RTE thresholds on the KITTI@30m and nuScenes@30m.

Model	KITTI@10m					KITTI@20m					KITTI@30m					KITTI@40m					
	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	RRE	RTE	RRE*	RTE*	RR	mRR
FCGF [8]	**0.196**	**0.060**	1.455	0.248	98.0	**0.345**	**0.105**	5.739	1.458	92.2	**0.701**	**0.219**	10.413	5.949	79.5	1.398	**0.350**	**17.755**	20.333	46.0	78.9
Predator [16]	0.415	0.093	1.086	0.254	97.8	1.267	0.249	9.003	7.430	59.4	1.881	0.279	19.460	28.709	0.5	–	–	26.632	38.956	0.0	39.4
CoFiNet [39]	0.345	0.084	0.345	0.090	**99.8**	0.939	0.201	3.355	1.541	90.0	1.843	0.610	13.620	17.622	31.9	2.181	1.183	26.385	36.950	4.3	56.5
GeoTrans [28]	0.550	0.130	8.157	6.938	51.4	0.894	0.315	16.391	8.515	45.2	1.593	0.557	44.138	21.729	13.5	1.974	0.847	43.806	32.264	10.8	30.2
BUFFER [2]	0.261	0.065	**0.264**	**0.072**	**99.8**	0.508	0.135	**0.539**	0.349	98.6	0.780	0.265	**4.873**	3.542	87.0	1.058	0.396	21.435	17.723	53.2	84.7
PARE [38]	0.557	0.160	16.274	4.878	57.3	1.739	0.700	45.434	18.671	2.1	–	–	54.498	29.199	0.0	–	–	65.780	39.359	0.0	14.9
GCL† [19]	0.265	0.078	0.680	0.129	99.5	0.448	0.116	5.169	1.311	93.2	0.742	0.173	9.694	4.626	83.8	1.262	0.309	19.348	13.943	63.3	85.0
UGP (Ours)	0.282	0.078	0.287	0.085	**99.8**	0.497	0.166	1.098	**0.309**	**98.9**	0.769	0.333	5.287	**1.867**	**90.8**	**0.975**	0.523	18.260	**8.791**	**74.1**	**90.9**

Table 2. **Cross-dataset generalization experiments.** We **train on the nuScenes@10m** dataset and then test on the KITTI@10m, KITTI@20m, KITTI@30m and KITTI@40m. '†' denotes that the method was trained using data within the range of [5m, 60m].

nuScenes. Experimental results are shown in Tab. 1. Our method surpasses all the above approaches in terms of RR on both the KITTI and nuScenes datasets, particularly outperforming BUFFER. Compared to BUFFER, our method improves KITTI mRR by 6.2% and KITTI@40m RR by 20.8%. Furthermore, on the much sparser nuScenes dataset, our method significantly outperforms all other methods. Our method achieves SOTA at nuScenes@10m and then at nuScenes@20m improves by 42.3% compared to GeoTrans, 14.7% compared to BUFFER and 15% compared to FCGF. At 30m our method still maintains 93.9%, and outperforms BUFFER by 47.3%. On the challenging 40m our method improves 57.1% compared to BUFFER.

We further analyze the influence of point cloud pair distance changes on different methods. For coarse-to-fine methods (CoFiNet, GeoTransformer and PARE) performance drops significantly as point cloud pair distance changes. Even when generalizing to the relatively simpler 20m data subset, GeoTransformer's RR drops to 7.5%, while PARE's RR drops to 1.8%, and their performance continues to degrade as the distance increases. In contrast, for the point-wise or patch-wise methods (FCGF and BUFFER) , which avoid inter-frame information exchange in their network design, retain their generalization capacity. When the cross attention is eliminated (CoFiNet.w/o C, GeoTrans.w/o C and PARE.w/o C in Tab. 1) , the generalization capability is unlocked. This further validates that the elimination of the cross-attention module plays a key role in enhancing the generalization of LiDAR scenes.

To evaluate the overall performance of our method, we also included average error metrics RRE* and RTE* for all test point pairs. Our method achieves the highest accuracy in terms of average RRE and RTE across all test point pairs for the 20m, 30m, and 40m distance generalization experiments. We display the cross-distance generalization registration recall with different RRE and RTE thresholds on two datasets in Fig. 5. According to this result, while the performance of BUFFER in KITTI@30m is higher than our method at some thresholds, we show a very large improvement at all thresholds in the larger nuScenes dataset test.

Model	nuScenes@10m				
	RRE(°)	RTE(m)	RRE*(°)	RTE*(m)	RR(%)
FCGF [8]	0.390	0.195	0.889	0.711	93.3
SpinNet [1]	1.941	0.390	10.714	7.562	7.1
Predator [16]	1.034	0.265	5.146	1.708	80.2
CoFiNet [39]	0.713	0.231	0.959	1.122	88.5
GeoTrans [28]	0.920	0.288	12.313	6.500	44.3
Buffer [2]	**0.367**	**0.183**	**0.435**	0.626	94.3
PARE [38]	1.371	0.402	7.901	6.489	52.5
UGP(Ours)	0.422	0.199	0.478	**0.322**	**98.4**

Table 3. The results of the cross-dataset generalization experiments **from KITTI@10m to nuScenes@10m**.

Cross-Datasets Generalization. To evaluate the cross-dataset generalization capability of the proposed UGP, we

No.	EC	PSA	BF	KITTI@30m PIR	IR	RRE	RTE	RRE*	RTE*	RR
a	-	-	-	1.3	1.5	1.712	0.531	37.897	33.874	6.5
b	✓	-	-	49.7	39.4	0.657	**0.280**	2.155	0.899	95.1
c	✓	✓	-	54.5	38.5	0.658	0.306	2.090	**0.541**	96.2
d	✓	-	✓	56.1	**41.5**	**0.619**	0.294	**2.032**	0.681	95.7
e	✓	✓	✓	**57.6**	40.6	0.677	0.290	2.127	0.805	**96.8**
No.	EC	PSA	BF	KITTI@40m PIR	IR	RRE	RTE	RRE*	RTE*	RR
a	-	-	-	0.8	0.7	2.101	0.967	54.430	46.728	2.2
b	✓	-	-	24.2	18.2	0.950	0.486	21.539	10.401	66.2
c	✓	✓	-	28.3	18.5	0.909	0.470	17.692	9.091	71.2
d	✓	-	✓	29.9	**20.3**	**0.907**	**0.461**	14.650	8.967	74.8
e	✓	✓	✓	**32.5**	19.9	0.975	0.494	**13.095**	**6.168**	**82.0**

Table 4. **Ablation experiment** analysis of UGP on KITTI. Eliminating the cross-attention module (EC), progressive self-attention module (PSA), BEV fusion (BF) are ablated.

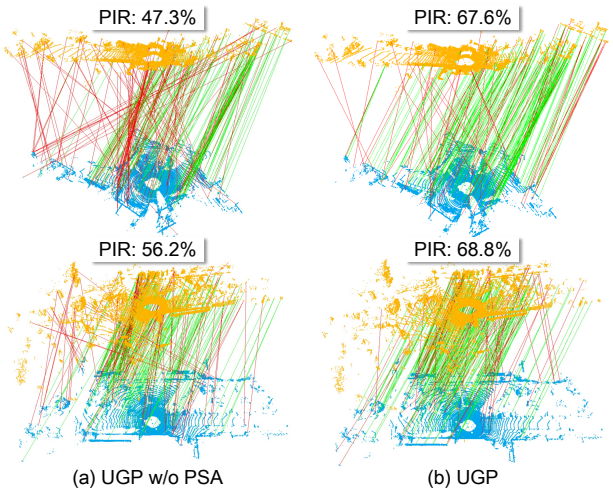


Figure 6. Registration results of the models with (a) vanilla self-attention and (b) progressive self-attention.

conducted the following experiments: 1) from nuScenes to KITTI@ (10m, 20m, 30m and 40m), 2) from KITTI to nuScenes. As shown in Tab. 2, the proposed UGP method, at KITTI@10m, achieved 99.8% registration recall. Also on KITTI@20m, KITTI@30m and KITTI@40m, the best results 98.9% (+0.3%), 90.8% (+3.8%) and 74.1% (+20.9%) were achieved compared to BUFFER. It can be found that our method has a significant advantage at longer distances. Finally, we compared the GCL [19] method using the pre-training weights of GCL in nuScenes. It can be found that our method also has a significant improvement in cross-dataset generalization compared to GCL, even though GCL utilizes [5m, 60m] point cloud frame during training. As shown in Tab. 3, our method achieves the best results despite the differences in sensor beams (KITTI: 64-line, nuScenes: 32-line). Compared to the sub-optimal BUFFER, the RR of our method improves by 4.1%.

5.3. Ablation Study

Ablation of Backbone. First, we performed an ablation study on the core components of the method to demonstrate the effectiveness of each part, as shown in Tab. 4. By comparing rows *a*-*b* in Tab. 4, we can see that EC significantly improves the network's generalization capability. By comparing rows *b*-*c*, we observe that PSA improves PIR by 4.8% and increases RR on KITTI@30m by 1.1% and by 5% at longer distances (40m) compared to the vanilla self-attention module. we further visulize the matching results of the ablation PSA, as shown in Fig. 6. It can be seen that the number of long-distance false matches is reduced significantly in the figure, proving that our PSA can effectively solve the feature ambiguity problem in vanilla self-attention. By comparing rows *b*-*d* and *c*-*e*, it can be observed that BEV provides semantic information about scene elements, reducing scene ambiguity. As a result, the framework's PIR and RR both improve. Column *e* in the table represents our complete proposed method, which achieves the best performance in the key metric of RR.

Estimators	mRR	KITTI@10m	KITTI@20m	KITTI@30m	KITTI@40m
UGP+RANSAC50K	92.4	**99.8**	**99.3**	95.1	75.5
UGP+MAC	**94.9**	**99.8**	**99.3**	96.2	**84.2**
UGP+LGR	94.5	**99.8**	**99.3**	**96.8**	82.0

Table 5. **Comparison of cross-distance generalization performance** for different estimators on KITTI.

Ablation of Estimator. We evaluated the impact of different estimators on the registration results, as shown in Tab. 5. The mRR of our method with RANSAC [11] is 92.4%, with MAC [42] is 94.9%, and with LGR [28] is 94.5%.

6. Conclusion

In this paper, we revealed that inconsistent geometric representations in LiDAR scenarios cause the cross-attention module to limit network generalization capability. Based on this insight, we proposed UGP, a pruned framework designed to enhance generalization power for LiDAR point cloud registration. UGP eliminated cross-attention, introduced a progressive self-attention module and a BEV feature extraction module, enabling the network to prioritize local spatial associations and captured semantic information of scene elements. This reduced ambiguity in point clouds and boosted generalization performance. Extensive experiments showed that our method effectively addresses challenges across varying data distributions, including cross-distance and cross-dataset scenarios.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (No.U23B2013, 62372150 and 62372377).

References

[1] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2021. 1, 2, 6, 7

[2] Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1264, 2023. 2, 6, 7, 3, 5

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020. 1, 2

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1, 6

[5] Anh-Quan Cao, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Pcam: Product of cross-attention matrices for rigid registration of point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13229–13238, 2021. 1

[6] Hong Chen, Pei Yan, Sihe Xiang, and Yihua Tan. Dynamic cues-assisted transformer for robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21698–21707, 2024. 2, 5, 6

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[8] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Cision*, pages 8958–8966, 2019. 2, 6, 7, 3

[9] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018. 1, 2

[10] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6433–6442, 2021. 3

[11] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 8

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 6, 5

[13] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 1, 2

[14] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11): 2270–2287, 2014. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 1

[16] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 1, 6, 7, 2

[17] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[18] Quan Liu, Yunsong Zhou, Hongzi Zhu, Shan Chang, and Minyi Guo. Apr: online distant point cloud registration through aggregated point cloud reconstruction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023. 2

[19] Quan Liu, Hongzi Zhu, Yunsong Zhou, Hongyang Li, Shan Chang, and Minyi Guo. Density-invariant features for distant point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18215–18225, 2023. 7, 8

[20] Quan Liu, Hongzi Zhu, Zhenxi Wang, Yunsong Zhou, Shan Chang, and Minyi Guo. Extend your own correspondences: Unsupervised distant point cloud registration by progressive distance extension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20816–20826, 2024. 2

[21] Zhe Liu, Shunbo Zhou, Chuanzhe Suo, Peng Yin, Wen Chen, Hesheng Wang, Haoang Li, and Yun-Hui Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2831–2840, 2019. 4

[22] Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei Li, and Hui-Liang Shen. Bevplace: Learning lidar-based place recognition using bird's eye view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8700–8709, 2023. 3

[23] Juncheng Mu, Lin Bie, Shaoyi Du, and Yue Gao. Colorpcr: Color point cloud registration with multi-stage geometric-color fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21061–21070, 2024. 2

[24] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cam-

eras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1

[25] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[27] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3979–3985, 2022. 1, 2

[28] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[29] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5135–5142. IEEE, 2020. 1

[30] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 1

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 2

[32] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019. 2, 4, 5, 1

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 4

[34] Kezheng Xiong, Maoji Zheng, Qingshan Xu, Chenglu Wen, Siqi Shen, and Cheng Wang. Speal: Skeletal prior embedded attention learning for cross-source point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6279–6287, 2024. 2

[35] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021. 3, 4

[36] Zongyi Xu, Xinqi Jiang, Xinyu Gao, Rui Gao, Changjun Gu, Qianni Zhang, Weisheng Li, and Xinbo Gao. Igreg: Image-geometry-assisted point cloud registration via selective correlation fusion. *IEEE Transactions on Multimedia*, 2024. 3

[37] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[38] Runzhao Yao, Shaoyi Du, Wenting Cui, Canhui Tang, and Chengwu Yang. Pare-net: Position-aware rotation-equivariant networks for robust point cloud registration. *arXiv preprint arXiv:2407.10142*, 2024. 1, 2, 3, 5, 6, 7

[39] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021. 1, 2, 3, 5, 6, 7

[40] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5384–5393, 2023. 1

[41] Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17702–17711, 2023. 2, 3, 5

[42] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2023. 8

[43] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. Pcr-cg: Point cloud registration via deep explicit color and geometry. In *Proceedings of the European Conference on Computer Vision*, pages 443–459. Springer, 2022. 3