PEAL: Prior-embedded Explicit Attention Learning for Low-overlap Point Cloud Registration

Junle Yu[1] Luwei Ren[1] Wenhui Zhou[1*] Yu Zhang[2*] Lili Lin[3] Guojun Dai[1]
[1]Hangzhou Dianzi University [2]Shanghai Jiaotong University [3]Zhejiang Gongshang University

Abstract

*Learning distinctive point-wise features is critical for low-overlap point cloud registration. Recently, it has achieved huge success in incorporating Transformer into point cloud feature representation, which usually adopts a self-attention module to learn intra-point-cloud features first, then utilizes a cross-attention module to perform feature exchange between input point clouds. The advantage of Transformer models mainly benefits from the use of self-attention to capture the global correlations in feature space. However, these global correlations may involve ambiguity for point cloud registration task, especially in indoor low-overlap scenarios, because the correlations with an extensive range of non-overlapping points may degrade the feature distinctiveness. To address this issue, we present PEAL, a **P**rior-embedded **E**xplicit **A**ttention **L**earning model. By incorporating prior knowledge into the learning process, the points are divided into two parts. One includes points lying in the putative overlapping region and the other includes points located in the putative non-overlapping region. Then PEAL explicitly learns one-way attention with the putative overlapping points. This simplistic design attains surprising performance, significantly relieving the aforementioned feature ambiguity. Our method improves the Registration Recall by 6+% on the challenging 3DLoMatch benchmark and achieves state-of-the-art performance on Feature Matching Recall, Inlier Ratio, and Registration Recall on both 3DMatch and 3DLoMatch.*

1. Introduction

Rigid point cloud registration has always been a foundational yet challenging task in 3D vision and robotics [2, 3, 10, 25], which aims to estimate an optimal rigid transformation to align two point clouds.

Benefiting from the superior feature representation of deep networks, keypoints-based point cloud registration methods have become dominant in recent years [4, 9, 12, 34, 37]. The core idea is to learn to match the learned keypoints across different point clouds. Recently, the keypoint-free methods [25, 36] demonstrate promising performance following the coarse-to-fine fashion. They seek correspondences between downsampled point clouds (superpoints), which are then propagated to individual points to yield dense correspondences. Thus, the accuracy of superpoint matching is crucial to the overall performance of point cloud registration. GeoTransformer [25] proposes a geometric self-attention module that encodes the distance of point pairs and the angle of triplet to extract transformation-invariant features. This approach significantly improves the accuracy of superpoint matching.

However, GeoTransformer may still suffer from ambiguous matching in certain scenarios with numerous similar structure or low geometric discriminative patches (superpoints) [25]. Moreover, the self-attention mechanism may exacerbate matching ambiguity, especially for low-overlap registration tasks. Prior works [3, 25] advocate that modeling geometric consistent correlations among overlap-

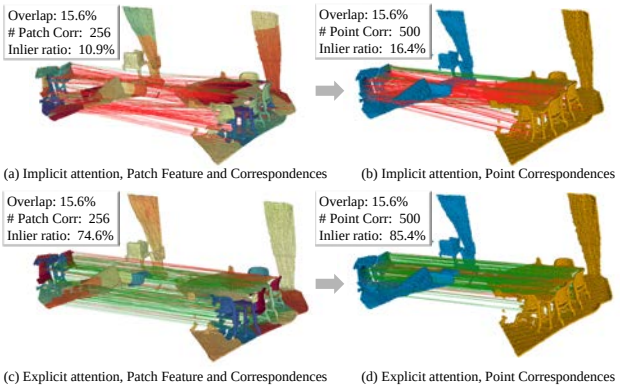


Figure 1. Given two low-overlap point clouds, PEAL adopts an explicit attention learning fashion and learns discriminative superpoint (patch) features (c), which results in significant higher patch and point inlier ratios. In contrast, GeoTransformer learns ambiguous patch features (a). For example, PEAL is able to accurately identify corresponding chairs among multiple chairs and distinguish them from the floor and table, while GeoTransformer mismatches them. Zoom in for details.

*Corresponding author: Wenhui Zhou, zhouwenhui@hdu.edu.cn; Yu Zhang, zhangyu606@gmail.com

ping superpoints/points is the key to the success of superpoints/points matching, while the global correlations learning via the geometric self-attention is inevitably interfered by numerous superpoints in the non-overlapping region. In other words, the correlations with non-overlapping superpoints may disrupt the inter-frame geometric consistency learning and degrade the feature distinctiveness for registration, which makes the resultant learned superpoint features ambiguous and leads to numerous outlier matches (Fig. 1 (a)).

To address the aforementioned issues, we design a **P**rior-embedded **E**xplicit **A**ttention **L**earning model (PEAL). It first leverages an overlap prior to divide the superpoints into anchor ones (the superpoints lying in putative overlapping region) and non-anchor ones (the superpoints located in putative non-overlapping region). Then it alleviates the interference of non-anchor superpoints by introducing an one-way attention mechanism, which solely models the correlations from non-anchor superpoints to anchor ones. Benefitting from the promising overlap ratio in the anchor region, anchor superpoints can be reckoned as simultaneously existing in both two frames, thus the one-way attention is capable of acquiring the essential local geometric consistent correlation from the anchor region, which helps the non-anchor superpoints encoding the inter-frame local geometric consistency and relieves the global feature ambiguity (Fig. 1 (c)). Furthermore, the embedding prior design involved in PEAL makes refining transformation possible in an iterative fashion.

In this paper, we introduce two models depending on the methods of obtaining prior, with extensive experiments on indoor benchmarks demonstrating the superiority of PEAL. Compared to state-of-the-art methods, both of the two models achieve significant improvements on Registration Recall on the challenging 3DLoMatch benchmark. In summary, our contributions are summarized as follows:

- To the best of our knowledge, we are the first to explicitly inject overlap prior into Transformer to facilitate low-overlap point cloud registration, and various overlap priors can be integrated into this framework, such as 3D overlap prior, 2D overlap prior, and self-overlap-prior.
- An explicit one-way attention module, which can significantly relieves the feature ambiguity generated by self-attention. It can be plugged into other transformer-based point cloud registration networks.
- A novel iterative pose refined fashion for low-overlap point cloud registration.

2. Related Work

Correspondence-based Point Cloud Registration. Correspondence-based methods [7, 9, 12] firstly establish correspondences between learned keypoints and then recover the transformation with a robust pose estimator such as RANSAC or other RANSAC-free estimator [3, 5, 6, 22, 25]. Many works focus on learning keypoint detectors [4, 17] and feature descriptors [1, 7, 32]. Recently, detector-free methods [25, 36] have become prevalent in a coarse-to-fine fashion. Our method inherits the detector-free methods and focuses on improving the matching accuracy of the coarse level.

2D-3D Multi-Modal Learning. Multimodal learning is currently a hot research area [8, 13, 15, 16, 19, 20] and 2D-3D joint learning is a typical multi-modal learning branch. Pri3d [15] employs an implicit pretrain-and-finetune strategy to combine 2D and 3D knowledge for 2D downstream tasks. 3D-SIS [13] and RevalNet [14] propose to implicitly fusing the color signal for 3D instance segmentation and detection tasks. ImVoteNet [23] explicitly uses 2D color input to boost 3D object detection. BYOC [11] and PCR-CG [38] explicitly fuse image features with point clouds to boost point cloud registration. Our method employs the explicit fashion to leverage the 2D signal.

Image matching. Image matching [21, 26, 28, 30] is a basic and important technology in computer vision. SIFT [21] and ORB [27] are typical handcrafted local features which are widely adopted in many 3D computer vision tasks. Learning-based methods [28, 30] can significantly improve the performance of local features under large viewpoints and illumination changes. We further explore how to leverage image signals to facilitate 3D point cloud registration.

Transformers. The Transformer models [31] utilize a novel attention mechanism, employing multiple layers of self and cross multi-head attention to facilitate information exchange between input and output, demonstrating the superior performance of feature representation for NLP and vision tasks. Focal self-attention [33] is to incorporate fine-grained local and coarse-grained global interactions to capture both short and long-range visual dependencies. Deformable DETR [39] uses deformable attention, which solely attends to a small set of key sampling points around a reference to aggregate multi-scale image features. GeoTransformer [25] utilizes geometric self-attention to extract transformation-invariant geometric features by encoding geometric structure. In this paper, we adopt an explicit attention learning fashion to learn distinctive features.

3. Method

We define point clouds $\mathcal{P} = \{\mathbf{p}_{x_i} \in \mathbb{R}^3 \mid i = 1, \dots, N\}$ and $\mathcal{Q} = \{\mathbf{q}_{y_i} \in \mathbb{R}^3 \mid i = 1, \dots, M\}$, as source and target, respectively. The purpose of point cloud registration is to recover the unknown rigid transformation $\mathrm{SE}(3)$, which consists of a rotation $\mathbf{R} \in \mathrm{SO}(3)$ and a translation $\mathrm{T} \in \mathbb{R}^3$ for aligning $\mathcal{P}$ and $\mathcal{Q}$.

The pipeline of our method is illustrated in Fig. 2. We

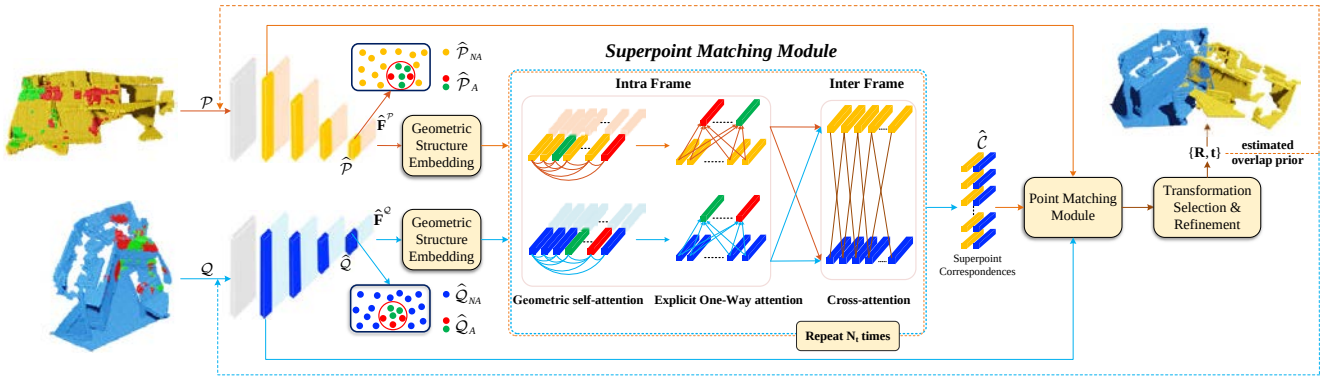
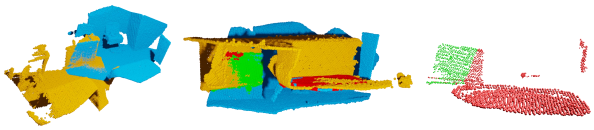


Figure 2. KPConv-FPN downsamples the input point clouds and learns features in multiple resolution levels. The geometric self-attention learns a hybrid geometric feature by encoding the geometric structure first, followed by intra-frame attention between anchor (selected according to the given overlap prior, painted in red or green) and non-anchor superpoints (painted in blue or yellow). This enables the learning of distinctive geometric features, which are then exchanged between two point clouds using a feature-based cross-attention. The resulting correspondences are then sent to the point matching module to calculate the transformation. During testing, the self-overlap-prior is computed based on the current transformation and iteratively transferred to the network.



(a) ground truth registration (b) GeoTransformer estimated registration (c) estimated overlap prior

Figure 3. The pipeline of acquiring 3D overlap prior. We show the correct estimated overlapping point cloud in green and the incorrect one in red.

first introduce how to acquire overlap prior (Sec. 3.1). Given the overlap prior, we then adopt the coarse-to-fine technique to extract correspondences. We build our method upon GeoTransformer [25], which utilizes KPConv-FPN to downsample the input point clouds and extract point-wise features simultaneously. The first and the last (coarsest) level of downsampled points correspond to the dense points and the superpoints. The superpoints are denoted by $\hat{\mathcal{P}}$, and $\hat{\mathcal{Q}}$, with the associated learned features denoted as $\hat{\mathbf{F}}^{\mathcal{P}}$ and $\hat{\mathbf{F}}^{\mathcal{Q}}$.

The superpoints with local patch containing overlap prior points are regarded as anchor superpoints, denoted as $\hat{\mathcal{P}_A}, \hat{\mathcal{Q}_A}$ for $\hat{\mathcal{P}}, \hat{\mathcal{Q}}$, respectively, while the other superpoints are viewed as non-anchor superpoints denoted as $\hat{\mathcal{P}}_{NA}, \hat{\mathcal{Q}}_{NA}$. And a superpoint matching module (Sec. 3.2) is used to extract superpoint correspondences. During superpoint matching, we first compute intra-frame geometric self-attention like GeoTransformer [25], then an explicit one-way attention module is proposed to encode the inter-frame local geometric consistency, which captures the one-way correlation from non-anchor superpoints to anchor ones. And finally, the inter-frame feature-based cross attention is computed to perform feature exchange between two input point clouds.

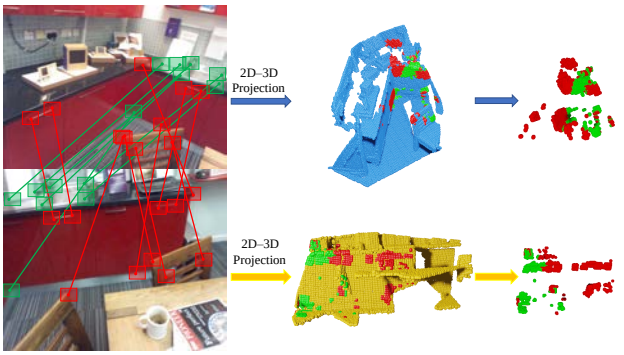
We follow GeoTransformer [25] for superpoint matching in order to extract the superpoint correspondences, which are then propagated to the point matching module to obtain dense point correspondences for calculating the final transformation. At last, the iterative update module (Sec. 3.3) is introduced, which can iteratively refine the transformation by using the predicted overlap prior (self-overlap-prior) generated by the currently estimated transformation.

3.1. Overlap Prior Prediction

Existing point cloud registration methods still face challenges in accurately registering two point clouds, especially in low-overlap scenarios. This often results in partially aligned or completely erroneous final transformations. Despite this, these partially aligned scenarios still show potential for achieving higher overlap. And the estimated overlapping regions in these cases can serve as a 3D overlap prior. Our goal is to improve these partially aligned scenarios by using the 3D overlap prior, transforming them into successfully registered ones.

On the other hand, inspired by multi-modal learning methods, it is worth considering the use of 2D images in RGB-D datasets. Image matching approaches could be utilized to estimate 2D correspondences, which can then be projected onto point clouds through a projection module. By doing so, we can obtain a preliminary estimation of the overlap between images, which serves as a 2D overlap prior.

3D overlap prior We use GeoTransformer [24] to generate 3D overlap prior, which is currently the state-of-the-art method. The predicted overlapping region is generated by the nearest neighbor search within a threshold in the Euclidean space according to the transformation matrix estimated by the pretrained GeoTransformer. The overlapping region is regarded as 3D overlap prior, namely 3dprior. The



(a) input images (b) point clouds with overlap prior (c) estimated overlap prior

Figure 4. The pipeline of acquiring 2D overlap prior. Similar to 3dprior, we show the correct estimated overlapping point cloud in green and the incorrect one in red.

pipeline of acquiring 3dprior is illustrated in Fig. 3.

2D overlap prior For the RGB-D dataset, we follow PCR-CG [38] which utilizes the 2D image matching method to generate 2D correspondences and convey them to point clouds via a 2D-3D projection module. Superglue [28] is utilized to extract image correspondences first. However, 2D correspondences are sparse, combined with the presence of invalid projections, making it difficult to obtain sufficient overlapping points. To tackle this issue, we follow PCR-CG [38] by creating a box for each matched pixel in every pair of image matches. Lastly, a projection module is utilized to lift 2D overlapping pixels to 3D overlapping points that serves as the overlap prior, called 2dprior. The pipeline of acquiring 2dprior is illustrated in Fig. 4.

3.2. Superpoint Matching

The registration performance of existing coarse to fine methods [25, 36] heavily relies on accurate superpoint matching, and unreliable superpoint matching causes the failure of the registration. Some recent methods, such as GeoTransformer [25], propose encoding global geometric structure to learn transformation-invariant features.

However, utilizing self-attention for learning intra-frame point-wise features may introduce ambiguity because self-attention learns the global correlations, in which the correlations with a large range of non-overlapping superpoints introduce ambiguous ones. While the correlations with the overlapping superpoints are the key to encoding inter-frame geometric consistency for the point cloud registration task.

We propose to model the correlation with overlapping superpoints to acquire inter-frame geometric consistent correlations. Specifically, we suggest incorporating overlap prior knowledge into the computation of point-wise features by modeling non-global correlations with the anchor superpoints. Compared to the original global overlap ratio, this approach allows for a higher overlap ratio in the anchor region and enables the network to avoid introducing ambiguous geometric correlations.

To this end, we propose an one-way attention module to explicitly learn one-way correlation with anchor superpoints. We achieve superpoint matching by utilizing three attention modules, including a geometric self-attention [25] module to learn intra-frame point-wise geometric features, an explicit one-way attention module to encode inter-frame local geometric consistency, and a feature-based cross-attention module to perform feature exchange between source and target point clouds. These three attention modules are interleaved for N_t times to extract hybrid features $\hat{\mathbf{H}}^{\mathcal{P}}$ and $\hat{\mathbf{H}}^{\mathcal{Q}}$ for reliable superpoint matching (see Fig. 2).

Geometric self-attention. During superpoint matching, given $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$, in order to learn point-wise attention features, we follow GeoTransformer [25] to compute the intra-frame self-attention $\mathbf{X}^{\hat{\mathcal{P}}}$ and $\mathbf{X}^{\hat{\mathcal{Q}}}$. Given the input feature matrix $\mathbf{X} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$, the output feature matrix $\mathbf{Z} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$ is obtained by performing a weighted sum of all projected input features:

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

the weight coefficient $a_{i,j}$ is obtained through a row-wise softmax function applied to the attention score $e_{i,j}$, and the computation of $e_{i,j}$ is shown as followed:

$$e_{i,j} = \frac{\left(\mathbf{x}_i \mathbf{W}^Q \right) \left(\mathbf{x}_j \mathbf{W}^K + \mathbf{r}_{i,j} \mathbf{W}^R \right)^T}{\sqrt{d_t}}. \tag{2}$$

Here, $\mathbf{r}_{i,j} \in \mathbb{R}^{d_t}$ is a geometric structure embedding which consists of a pair-wise distance embedding and a triplet-wise angular embedding [25], $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \mathbf{W}^R \in \mathbb{R}^{d_t \times d_t}$ are projection matrices for queries, keys, values, and geometric structure embeddings, respectively. Then, we get superpoints attention feature matrices $\mathbf{X}^{\hat{\mathcal{P}}}$ and $\mathbf{X}^{\hat{\mathcal{Q}}}$, anchor superpoints attention feature matrices $\mathbf{X}^{\hat{\mathcal{P}}_A}$ and $\mathbf{X}^{\hat{\mathcal{Q}}_A}$. $\mathbf{X}^{\hat{\mathcal{P}}_A}$ represents the anchor superpoints attention matrices of $\hat{\mathcal{P}}$, which is indexed from the self-attention matrices $\mathbf{X}^{\hat{\mathcal{P}}}$ according to divided putative overlapping superpoints, and the same goes for $\mathbf{X}^{\hat{\mathcal{Q}}_A}$. Similarly, $\mathbf{X}^{\hat{\mathcal{P}}_{NA}}$ and $\mathbf{X}^{\hat{\mathcal{Q}}_{NA}}$ represent the non-anchor superpoints attention feature matrices for $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$, respectively.

Explicit one-way attention. We propose an one-way attention module to explicitly learn the intra-frame correlation with the anchor superpoints, which is key to encoding inter-frame local geometric consistency. Given anchor superpoints attention feature matrices $\mathbf{X}^{\hat{\mathcal{P}}_A}$ and non-anchor superpoints attention feature matrices $\mathbf{X}^{\hat{\mathcal{P}}_{NA}}$, the non-anchor superpoints attention feature matrices $\mathbf{Z}^{\hat{\mathcal{P}}_{NA}}$ can be com-

puted with the features of anchor superpoints features $\mathbf{X}^{\hat{\mathcal{P}}_A}$.

$$\mathbf{z}_m^{\hat{\mathcal{P}}_{NA}} = \sum_{n=1}^{|\hat{\mathcal{P}}_A|} a_{m,n} \left(\mathbf{x}_n^{\hat{\mathcal{P}}_A} \mathbf{W}^V \right) \tag{3}$$

Similar to geometric self-attention, $a_{m,n}$ represents a row-wise softmax on the attention score $e_{m,n}$, which is the feature correlation between the $\mathbf{X}^{\hat{\mathcal{P}}_{NA}}$ and $\mathbf{X}^{\hat{\mathcal{P}}_A}$.

$$e_{m,n} = \frac{\left(\mathbf{x}_m^{\hat{\mathcal{P}}_{NA}} \mathbf{W}^{\hat{\mathcal{P}}_A} \right) \left(\mathbf{x}_n^{\hat{\mathcal{P}}_A} \mathbf{W}^K \right)^T}{\sqrt{d_t}} \tag{4}$$

The attention features for $\mathbf{X}^{\hat{\mathcal{Q}}_{NA}}$ are updated same as $\mathbf{X}^{\hat{\mathcal{P}}_{NA}}$, while $\mathbf{X}^{\hat{\mathcal{Q}}_A}$ and $\mathbf{X}^{\hat{\mathcal{P}}_A}$ remain unchanged.

Next, we use an inter-frame feature-based cross-attention module to encode global inter-frame geometric consistency between input point clouds [25]. The resultant superpoint features are then sent to the superpoint matching module for reasoning precise superpoint correspondences. Finally, we perform point matching [25] to extract the dense point correspondences and local-to-global registration to obtain the estimated transformation.

3.3. Iterative Update

Our superpoint matching module incorporates an explicit one-way attention module that is sensitive to anchor superpoints, allowing for better-estimated transformations by improved overlap ratio in the anchor region. This leads to more distinctive features for superpoint matching and reliable correspondences, resulting in higher accuracy and more reliable final transformation. In other words, it can be designed in an iterative fashion to refine the transformation. The iterative update module starts from the currently estimated transformation, and the self-overlap-prior is computed according to the current transformation and iteratively transferred into the network. At each iteration, it produces an updated transformation T_k, which is then applied to estimate $T_{k+1} = f(T_k)$.

4. Experimental Results

In this section, we evaluate our method on indoor 3DMatch [37] and 3DLoMatch [17] benchmarks.

4.1. 3DMatch & 3DLoMatch

Dataset. 3DMatch [37] is composed of 62 scenes among which 46 are used for training, 8 for validation, and 8 for testing. Each scene has its corresponding RGB-D data. We evaluate our approach using preprocessed training point clouds provided by [17] and test it on both the 3DMatch and 3DLoMatch protocols. The point cloud pairs in 3DMatch have an overlap of more than 0.3, while those in 3DLoMatch have a lower overlap of $0.1 \sim 0.3$. We collect the associated RGB-D data [38] of these two benchmarks, where each point cloud is fused by 50 consecutive depth frames.

	3DMatch					3DLoMatch				
Samples	5000	2500	1000	500	250	5000	2500	1000	500	250
Feature Matching Recall(%) ↑										
PerfectMatch [12]	95.0	94.3	92.9	90.1	82.9	63.6	61.7	53.6	45.2	34.2
FCGF [7]	97.4	97.3	97.0	96.7	96.6	76.6	75.4	74.2	71.7	67.3
D3Feat [4]	95.6	95.4	94.5	94.1	93.1	67.3	66.7	67.0	66.7	66.5
SpinNet [1]	97.6	97.2	96.8	95.5	94.3	75.3	74.9	72.5	70.0	63.6
Predator [17]	96.6	96.6	96.5	96.3	96.5	78.6	77.4	76.3	75.7	75.3
YOHO [32]	98.2	97.6	97.5	97.7	96.0	79.4	78.1	76.3	73.8	69.1
CoFiNet [36]	98.1	98.3	98.1	98.2	98.3	83.1	83.5	83.3	83.1	82.6
PCR-CG [38]	97.4	97.5	97.7	97.3	97.6	80.4	82.2	82.6	83.2	82.8
GeoTransformer [25]	97.9	97.9	97.9	97.9	97.6	88.3	88.6	88.8	88.6	88.3
PEAL(ours)	**99.0**	**99.0**	**99.1**	**99.1**	**98.8**	**91.7**	**92.4**	**92.5**	**92.9**	**92.7**
Inlier Ratio (%) ↑										
PerfectMatch [12]	36.0	32.5	26.4	21.5	16.4	11.4	10.1	8.0	6.4	4.8
FCGF [7]	56.8	54.1	48.7	42.5	34.1	21.4	20.0	17.2	14.8	11.6
D3Feat [4]	39.0	38.8	40.4	41.5	41.8	13.2	13.1	14.0	14.6	15.0
SpinNet [1]	47.5	44.7	39.4	33.9	27.6	20.5	19.0	16.3	13.8	11.1
Predator [17]	58.0	58.4	57.1	54.1	49.3	26.7	28.1	28.3	27.5	25.8
YOHO [32]	64.4	60.7	55.7	46.4	41.2	25.9	23.3	22.6	18.2	15.0
CoFiNet [36]	49.8	51.2	51.9	52.2	52.2	24.4	25.9	26.7	26.8	26.9
GeoTransformer [25]	71.9	75.2	76.0	82.2	85.1	43.5	45.3	46.2	52.9	57.7
PEAL(ours)	**72.4**	**79.1**	**84.1**	**86.1**	**87.3**	**45.0**	**50.9**	**57.4**	**60.3**	**62.2**
Registration Recall(%) ↑										
PerfectMatch [12]	78.4	76.2	71.4	67.6	50.8	33.0	29.0	23.3	17.0	11.0
FCGF [7]	85.1	84.7	83.3	81.6	71.4	40.1	41.7	38.2	35.4	26.8
D3Feat [4]	81.6	84.5	83.4	82.4	77.9	37.2	42.7	46.9	43.8	39.1
SpinNet [1]	88.6	86.6	85.5	83.5	70.2	59.8	54.9	48.3	39.8	26.8
Predatorr [17]	89.0	89.9	90.6	88.5	86.6	59.8	61.2	62.4	60.8	58.1
YOHO [32]	90.8	90.3	89.1	88.6	84.5	65.2	65.5	63.2	56.5	48.0
CoFiNet [36]	89.3	88.9	88.4	87.4	87.0	67.5	66.2	64.2	63.1	61.0
PCR-CG [38]	89.4	90.7	90.0	88.7	86.8	66.3	67.2	69.0	68.5	65.0
GeoTransformer [25]	92.0	91.8	91.8	91.4	91.2	75.0	74.8	74.2	74.1	73.5
PEAL(ours)	**94.6**	**93.7**	**93.7**	**93.9**	**93.4**	**81.7**	**81.2**	**80.8**	**80.4**	**80.1**

Table 1. Evaluation results on 3DMatch and 3DLoMatch.

The RGB images and depth images are in pairs. Therefore, each point cloud is also associated with 50 RGB frames.

Experiments Setup. When acquiring the 3D overlap prior, the predicted overlap region is generated by the nearest neighbor search under the estimated transformation within a threshold of 0.0375m. To obtain 2D overlap prior, considering the computation cost of inferencing 2D correspondences, we use three frames of images for the source and the target point clouds and extract correspondences from nine pairs of images using Superglue [28], with an image resolution of [640,480]. We project the XYZ coordinates of each point cloud onto its associated image plane to obtain pixel-point corresponding relation with a depth threshold of 0.2m. During training, we use an Adam optimizer with a learning rate of 1e-4 and a batch size of 1 on a single GPU (RTX3090) for a total of 20 epochs. For testing, we always use PEAL-3dprior for iterative updates and perform six iterations to refine the transformation. Other experimental setup goes the same with GeoTransformer [25].

Metrics. We mainly compare the results on five metrics as prior works [17, 25, 38], namely Registration Recall (RR), Feature Matching Recall (FMR), Inlier ratio (IR), Relative Rotation Error (RRE) as well as Relative Translation Error (RTE).

We compare our method with the recent state of the arts: FCGF [7], D3Feat [4], SpinNet [1], Predator [17], YOHO [32], CoFiNet [36], PCR-CG [38], GeoTransformer [25] ,

Model	Estimator	Samples	RR(%) 3DM	3DLM
FCGF [7]	RANSAC-*50k*	5000	85.1	40.1
D3Feat [4]	RANSAC-*50k*	5000	81.6	37.2
SpinNet [1]	RANSAC-*50k*	5000	88.6	59.8
Predator [17]	RANSAC-*50k*	5000	89.0	59.8
PCR-CG [38]	RANSAC-*50k*	5000	89.4	66.3
CoFiNet [36]	RANSAC-*50k*	5000	89.3	67.5
Lepard [18]	RANSAC-*50k*	5000	93.5	69.0
GeoTransformer [25]	RANSAC-*50k*	5000	92.0	75.0
PEAL-3dprior(ours)	RANSAC-*50k*	5000	94.4	79.4
PEAL-2dprior(ours)	RANSAC-*50k*	5000	**94.6**	**81.7**
CoFiNet [36]	LGR	all	87.6	64.8
REGTR [35]	RANSAC-free	all	92.0	64.8
GeoTransformer [25]	LGR	all	91.5	74.0
PEAL-3dprior(ours)	LGR	all	94.1	78.8
PEAL-2dprior(ours)	LGR	all	**94.3**	**81.2**

Table 2. Registration results w/o RANSAC on 3DMatch (3DM) and 3DLoMatch (3DLM).

REGTR [35], and Lepard [18].

RANSAC estimator results. Following [4, 17], we report the results with different numbers of correspondences using the RANSAC estimator in Tab. 1. Our method achieves the highest *Feature Matching Recall* on all the sampled correspondences on 3DMatch and 3DLoMatch. Our method improves by no less than 3.4% on 3DLoMatch and 1.1 % on 3DMatch compared to the baseline GeoTransformer [25]. For *Inlier Ratio*, PEAL improves by $1.5\% \sim 11.2\%$ on 3DLoMatch, and improves by $0.5\% \sim 8.1\%$ on 3DMatch. The improvement is more prominent with fewer sampled correspondences, for instance, we achieve 60+% *Inlier Ratio* on 500 and 250 sampled correspondences on 3DLoMatch. For *Registration Recall* , PEAL outperforms the previous best (Lepard [18] in 3DMatch Tab. 2 (top), GeoTransformer [25] in 3DLoMatch Tab. 1 (bottom)) by 1.1% on 3DMatch and 6.7% on 3DLoMatch.

RANSAC-free estimator. We compare the registration results of the RANSAC-free estimator in Tab. 2 (bottom). We separate our models into PEAL-2dprior and PEAL-3dprior according to the method of acquiring overlap prior. Both of them achieve state-of-the-art performance on 3DMatch 3DLoMatch. The improvement is more prominent than using the RANSAC estimator. PEAL-2dprior improves previous best (REGTR [35] in 3DMatch, GeoTransformer [25] in 3DLoMatch) by 2.3% on 3DMatch and 7.2% on 3DLoMatch. In addition, PEAL-3dprior outperforms the baseline by 2.6% on 3DMatch and 4.8% on 3DLoMatch, showing significant improvement compared to baseline method.

RRE and RTE. We then compare the RRE and RTE with the recent state of the arts: [1, 4, 7, 17, 25, 35, 36] in Tab. 3. As shown in this table, we achieve the second best in RRE, RTE on 3DMatch, and RTE on 3DLoMatch, and the best in RRE on 3DLoMatch.

	Estimator	3DMatch RRE (°)	3DMatch RTE (m)	3DLoMatch RRE (°)	3DLoMatch RTE (m)
Predator [17]	RANSAC-*50k*	2.029	0.064	3.048	0.093
CoFiNet [36]	RANSAC-*50k*	2.002	0.064	3.271	0.090
PCR-CG [38]	RANSAC-*50k*	1.993	0.061	3.002	0.087
Geotransformer [25]	RANSAC-free	1.772	0.061	2.849	0.088
REGTR [35]	RANSAC-free	**1.567**	**0.049**	2.827	**0.077**
PEAL-3dprior(ours)	RANSAC-free	1.745	0.061	2.802	0.087
PEAL-2dprior(ours)	RANSAC-free	1.748	0.062	**2.788**	0.087

Table 3. Relative Rotation Errors (RRE) and Relative Translation Errors (RTE) on 3DMatch and 3DLoMatch benchmarks.

one-way attention	3DMatch FMR	3DMatch IR	3DMatch RR	3DLoMatch FMR	3DLoMatch IR	3DLoMatch RR
none	97.7	70.3	91.5	88.1	43.3	74.0
non-anchor to anchor	**98.7**	**73.4**	**94.1**	**88.8**	**46.7**	**78.8**
anchor to non-anchor	63.7	33.1	54.9	22.0	6.3	16.8
bidirectional	98.1	63.3	90.9	85.0	36.4	71.6

Table 4. Ablation experiments about the explicit one-way attention module using PEAL-3dprior.

Ablation studies. In this section, we ablate the different setups for the proposed modules in our method. In practice, we notice local-to-global registration (LGR) [25] estimator is much more stable and faster than the RANSAC estimator while achieving a comparable performance (see Tab. 2). Thus, we conduct our ablation experiments based on the LGR estimator.

To explore the effectiveness of explicit one-way attention module, we conduct various one-way attention strategies between anchor and non-anchor superpoints, reporting the FMR, IR, and RR on 3DMatch and 3DLoMatch. We compared three strategies: anchor to non-anchor, non-anchor to anchor, and bidirectional (combine both non-anchor to anchor and anchor to non-anchor). Our findings indicate that the local geometry correlation from anchor points to non-anchor points is inter-frame geometrically inconsistent when performing anchor to non-anchor attention. When updating the features of anchor points, they will be severely contaminated by the non-anchor region, leading to a significant drop in registration recall compared to GeoTransformer, as shown in Tab. 4. Overall, our study confirms that explicit one-way attention from non-anchor superpoints to anchor superpoints can effectively encode inter-frame local geometric consistency.

To evaluate prior-embedded explicit attention learning, we compare it with a prior-embedded implicit attention learning approach that involves making anchor points salient. Specifically, prior to computing the self-attention, the anchor points are initialized with the pretrained KPConv

Model	3DMatch			3DLoMatch		
	FMR	IR	RR	FMR	IR	RR
GeoTransformer [25]	97.7	70.3	91.5	88.1	43.3	74.0
PIAL-3dprior	97.6	70.4	91.7	87.9	43.5	74.7
PEAL-3dprior	**98.7**	**73.4**	**94.1**	**88.8**	**46.7**	**78.8**

Table 5. Ablation experiments about the implicit attention learning and explicit attention learning with embedded prior.

Model	anchor superpoints selection	RR(%)	
		3DM	3DLM
PEAL-3dprior	random	91.5	73.8
PEAL-3dprior	random choose 40%	93.5	78.0
PEAL-3dprior	random choose 80%	94.1	78.2
PEAL-3dprior	base region	**94.1**	78.8
PEAL-3dprior	+ 4 nearest superpoints	93.9	**78.9**
PEAL-3dprior	+ 12 nearest superpoints	93.5	77.4
PEAL-3dprior	+ 16 nearest superpoints	93.2	77.2

Table 6. Ablation experiments on the anchor region. We take the overlapping region estimated by GeoTransformer [25] as the base region.

Model	Views	RR(%)	
		3DM	3DLM
PEAL-2dprior	1	92.5	78.8
PEAL-2dprior	2	93.7	80.7
PEAL-2dprior	3	**94.3**	**81.2**
PEAL-2dprior	4	94.3	81.2

Table 7. Ablation experiments on the number of images.

features while the non-anchor points are initialized to a constant. We report the FMR, IR and RR of prior-embedded implicit attention learning (PIAL) and prior-embedded explicit attention learning (PEAL) using LGR estimator in Tab. 5. When using 3dprior, PIAL-3dprior have not shown obvious improvements, while PEAL-3dprior significantly outperforms the baseline, providing further evidence of the superiority of the proposed approach.

Further, we vary the anchor region to study the influence on the prior overlapping region. In 3dprior experiments, we use the overlapping region estimated by GeoTransformer [25] as the base region, which is then expanded by including the nearest N superpoints or reduced by randomly selecting a certain percentage from the base region. The registration recall for 4, 12, and 16 nearest superpoints and 40% and 80% shrinkage on the base region is recorded in Tab. 6. For the 2dprior experiments, we adjust the number of images from 1-4 to obtain the correspondences and record the registration recall in Tab. 7.

Tab. 6 explores the registration performance of using random anchor superpoints first, as shown in the first row.

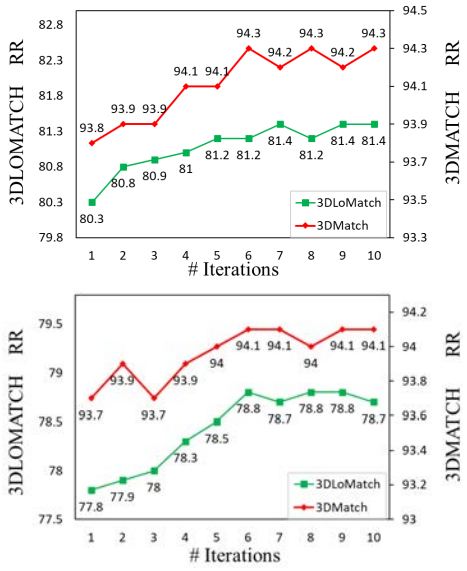


Figure 5. Registration Recall of iterative PEAL-3dprior (bottom) and iterative PEAL-2dprior (top).

Model	Iteration	RR(%)	
		3DM	3DLM
PEAL-3dprior	w/o iteration	93.7	77.8
PEAL-3dprior-iter	6 iteration	94.1	78.8
ICP [29]	20 iteration	94.1	78.2
PEAL-2dprior	w/o iteration	93.8	80.3
PEAL-2dprior-iter	6 iteration	94.3	81.2
ICP [29]	30 iteration	94.4	80.9

Table 8. Ablation experiments on the iterative update module.

The performance is close to baseline, implying that our method does not perform well with randomly initialized overlap prior. However, randomly expanding/shrinking the base anchor region by 20% or choosing the 16 nearest nodes only results in a slight drop in performance, suggesting that changes to the base anchor region have limited effect on the overlap ratio and consequently do not significantly affect the final performance.

Tab. 7 demonstrates that registration recall can be improved with more views involved, because more overlapping point clouds can be found with the number of views growing, especially in low-overlap scenarios with large viewpoints and illumination changes.

Next, we ablate the iterative update module in Tab. 8 and Fig. 5. We evaluate the number of iterations on registration recall. We use the pretrained PEAL to generate the initial pose for ICP and PEAL-iter. The registration recall continues to improve with increasing iterations, as depicted in Fig. 5. The improvement slows down after 8 iterations, so we set the iteration number to 6 in our experiments tak-

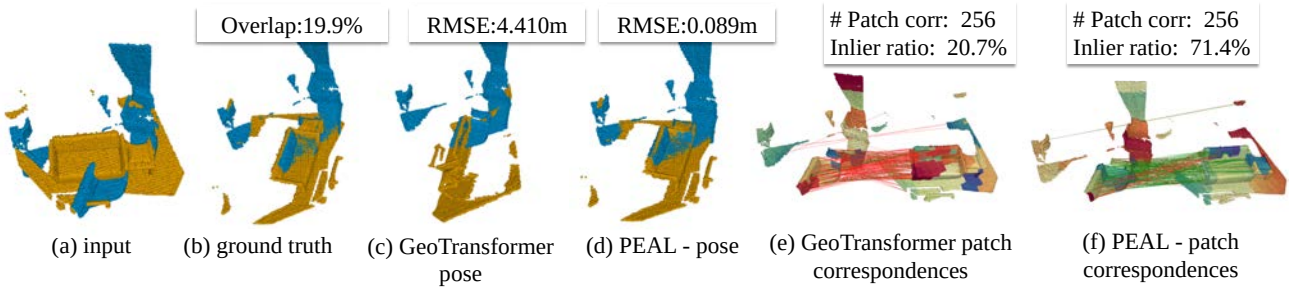


Figure 6. Registration results of the GeoTransformer and PEAL-3dprior. The overlapping region estimated by GeoTransformer is used as our 3D overlap prior. Our method infers much more inlier superpoint matches, significantly improving the registration performance. We use T-SNE to visualize superpoint (patch) features.

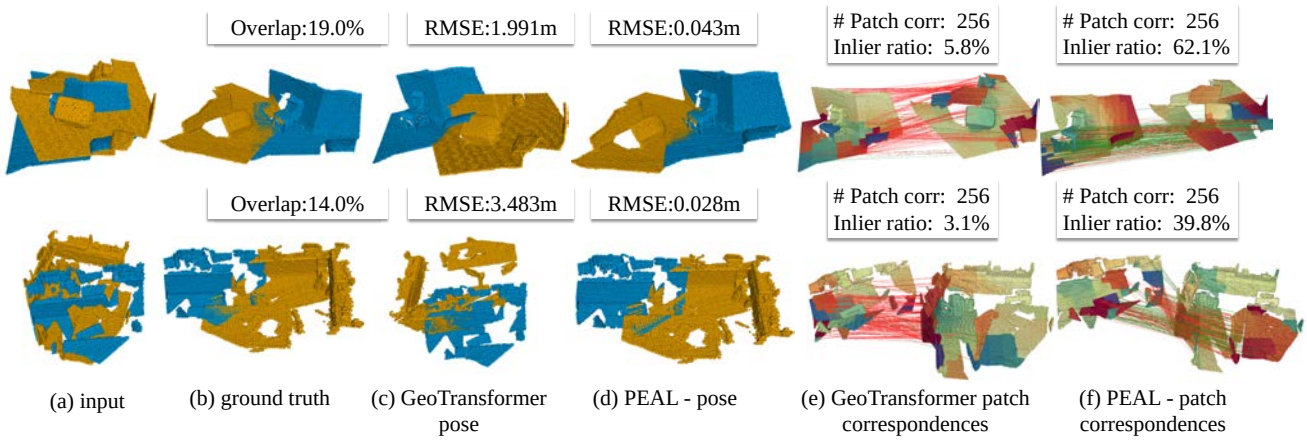


Figure 7. Registration results of GeoTransformer and PEAL-2dprior. Benefitting from the powerful overlap prior from 2D, PEAL-2dprior learns distinctive patch features and infers reliable patch matches on the structure-less floor and geometrically ambiguous cupboard.

ing the tradeoff between performance and computation cost. We also provide the RR of ICP in Tab. 8 and discover that ICP's performance is highly dependent on iteration number and distance thresholds. We exhaustively adjust the setting to identify the highest RR. Obviously, our approach demonstrates better performance in low-overlap scenarios, while ICP's performance does not continually increase with an increasing number of iterations.

Qualitative results. We visualize a collection of the registration results of PEAL and GeoTransformer [25] in Fig. 6, 7 . Our explicit attention-learning fashion significantly improves feature representation and helps to infer superpoint (patch) matches in extreme low-overlap scenarios.

Combining with other transformer-based networks. In Tab. 9, we combine PEAL with other transformer-based point cloud registration methods to evaluate its performance. When plugged into Predator [17], which is not a coarse-to-fine method, the improvement is also prominent. Specifically, the RR improves by 2.8% with 3D overlap prior, and improves by 10.4% with 2D overlap prior on 5000 sampled points.

	Registration Recall (%)				
# Sampled Points	5000	2500	1000	500	250
Predator [17]	59.8	61.2	62.4	60.8	58.1
Predator+PEAL	**70.2**	**70.2**	**69.4**	**68.6**	**63.3**

Table 9. Compared with Predator baseline on 3DLoMatch.

5. Conclusion

We present PEAL for rigid point cloud registration, which explicitly incorporates overlap prior to attention learning. It yields distinctive features for superpoint and dense point matching, and various overlap priors can be integrated into this framework, such as 3D overlap prior, 2D overlap prior, and self-overlap-prior. PEAL learns to extract the discriminative geometric features through explicit one-way attention, significantly relieving the feature ambiguity generated by self-attention.

Acknowledgments This work is supported in part by the National Key R&D Program of China (2017YFE0118200), the Zhejiang Provincial Natural Science Foundation of China (LTY22F020001, LY21F010007) and the National Natural Science Foundation of China (62076083).

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 *CVPR*, pages 11753–11762, 2021. 2, 5, 6

[2] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *CVPR*, pages 7163–7172, 2019. 1

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *CVPR*, pages 15859–15869, 2021. 1, 2

[4] 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 *CVPR*, pages 6359–6367, 2020. 1, 2, 5, 6

[5] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *CVPR*, pages 13221–13231, 2022. 2

[6] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, pages 2514–2523, 2020. 2

[7] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *CVPR*, pages 8958–8966, 2019. 2, 5, 6

[8] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *ECCV*, pages 452–468, 2018. 2

[9] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3D local descriptors. In *ECCV*, 2018. 1, 2

[10] Mohamed El Banani, Luya Gao, and Justin Johnson. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *CVPR*, pages 7129–7139, 2021. 1

[11] Mohamed El Banani and Justin Johnson. Bootstrap Your Own Correspondences. In *ICCV*, 2021. 2

[12] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *CVPR*, pages 5545–5554, 2019. 1, 2, 5

[13] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *CVPR*, 2019. 2

[14] Ji Hou, Angela Dai, and Matthias Nießner. RevealNet: Seeing Behind Objects in RGB-D Scans. In *CVPR*, 2020. 2

[15] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? *arXiv preprint arXiv:2104.11225*, 2021. 2

[16] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, pages 14373–14382, 2021. 2

[17] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *CVPR*, pages 4267–4276, June 2021. 2, 5, 6, 8

[18] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *CVPR*, pages 5554–5564, 2022. 6

[19] Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas Funkhouser, and Hao Dong. P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. *arXiv preprint arXiv:2012.13089*, 2020. 2

[20] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3d-to-2d distillation for indoor scene parsing. In *CVPR*, pages 4464–4474, 2021. 2

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[22] G Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C Nascimento, Rama Chellappa, and Pedro Miraldo. 3dregnet: A deep neural network for 3d point registration. In *CVPR*, pages 7193–7203, 2020. 2

[23] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3D object detection in point clouds with image votes. In *CVPR*, 2020. 2

[24] Charles Ruizhongtai Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas Guibas. Volumetric and multi-view cnns for object classification on 3D data. In *CVPR*, 2016. 3

[25] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, pages 11143–11152, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[26] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 2

[27] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 2

[28] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 2, 4, 5

[29] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 7

[30] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[32] Haiping Wang, Yuan Liu, Zhen Dong, and Wenping Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *MM*, pages 1630–1641, 2022. 2, 5

[33] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2

[34] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *ECCV*, pages 607–623, 2018. 1

[35] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *CVPR*, pages 6677–6686, 2022. 6

[36] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *NeurIPS*, 34:23872–23884, 2021. 1, 2, 4, 5, 6

[37] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *CVPR*, 2017. 1, 5

[38] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. Pcr-cg: Point cloud registration via deep explicit color and geometry. In *ECCV*, pages 443–459. Springer, 2022. 2, 4, 5, 6

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2