Proceedings Track

BiEquiFormer: Bi-Equivariant Representations for Global Point Cloud Registration

Editors: Congyue Deng, Maksim Zhdanov, Manos Theodosis

Abstract

The goal of this paper is to address the problem of *global* point cloud registration (PCR) i.e., finding the optimal alignment between point clouds irrespective of the initial poses of the scans. This problem is notoriously challenging for classical optimization methods due to computational constraints. First, we show that many state-of-the-art deep learning methods suffer from huge performance degradation when the point clouds are arbitrarily placed in space. We propose that *equivariant deep learning* should be utilized for solving this task and we characterize the specific type of bi-equivariance of PCR. Then, we design BiEquiformer a novel and scalable *bi-equivariant* pipeline i.e. equivariant to the independent transformations of the input point clouds. While a naive approach would process the point clouds independently we design expressive bi-equivariant layers that fuse the information from both point clouds. This allows us to extract high-quality superpoint correspondences and in turn, robust point-cloud registration. Extensive comparisons against state-of-the-art methods show that our method achieves comparable performance in the canonical setting and superior performance in the robust setting in both the 3DMatch and the challenging low-overlap 3DLoMatch dataset.

Keywords: Equivariance, Bi-Equivariance, Point-Cloud Registration, Geometric Deep Learning

1. Introduction

Point Cloud Registration (PCR) is at the frontend of many robotics and vision pipelines. The goal, in the pairwise and rigid setting, is to align two partially overlapped point clouds expressed in their own coordinate system by estimating a roto-translation between them and fusing them in a common coordinate system. It has been successfully applied in many tasks such as 3D Scene Reconstruction Blais and Levine (1995), SLAM (Nüchter et al., 2006) and pose estimation Yang et al. (2013).

While PCR has been studied extensively over the past decades, the desiderata for real-time and robust registration of real-world applications makes the problem extremely challenging. Especially in environments with repetitive patterns such as indoor environments as well as in low-overlap settings that appear loop closure tasks Bosse and Zlot (2008) the requirement for distinctive point-wise features for correspondence is enhanced. A particularly challenging aspect of the problem is the robustness w.r.t. the initial poses of the point clouds. In classical optimization methods, the problem is called *global* PCR and is famously intractable due to the large volume of points Yang et al. (2013).

Deep learning has been proven very effective in PCR in all building blocks of the registration pipeline. Powerful point cloud architectures Qi et al. (2016); Thomas et al. (2019) serve both as the feature extraction for correspondence-based methods Zeng et al. (2017); Choy et al. (2019) and a way to identify distinctive features for matching Huang et al. (2020); Li

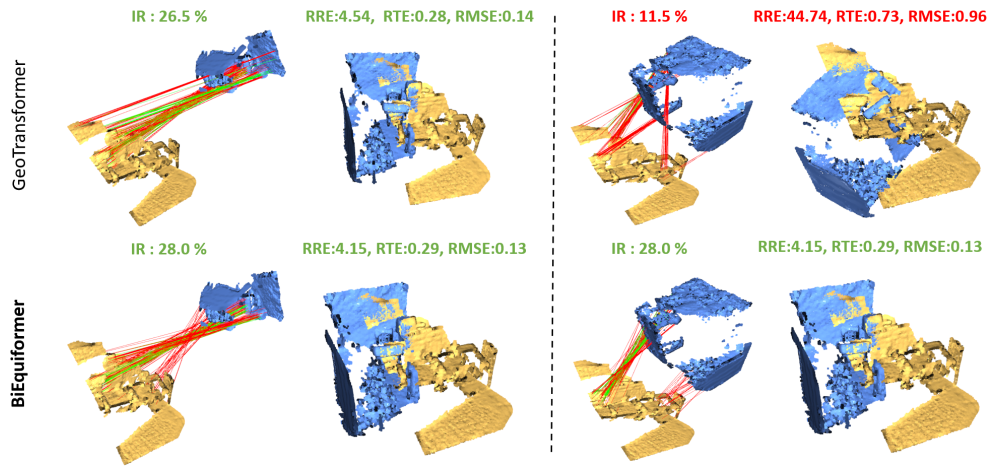

Figure 1: Registration Metrics (IR,RRE,RTE,RMSE) for two pairs of low-overlap scans that differ only by their relative pose. While both methods succeed in the original point cloud input (left column), GeoTransformer fails to find good matches (low IR) when the inputs are randomly rotated (right column), while our proposed BiEquiformer performs consistently irrespective of the initial point cloud poses.

and Harada (2022). It has also been utilized to learn robust estimators Choy et al. (2020); Pais et al. (2019); Bai et al. (2021) or directly regress the relative transformation (Wang and Solomon, 2019; Aoki et al., 2019). In this work, we show how recent state-of-the-art registration pipelines are heavily affected by the orientations of the initial scans, especially in challenging low-overlap settings (Fig. 3). Subsequently, we propose BiEquiformer a detector-free attention pipeline that is bi-equivariant to the roto-translation group (Fig.2). Our main contributions can be summarized as follows:

1. **The state of Global PCR in DL**: We investigate the robustness of state-of-the-art methods under rigid transformations of the input point clouds. In Fig. 3 we show that in numerous popular state-of-the-art methods there is a deterioration in performance when the initial poses of the point clouds vary, exacerbated as the overlap between scans becomes smaller. Figure 1 shows a visual example of this phenomenon.
2. **Bi-Equivariance and PCR**: We formulate and characterize the specific bi-equivariance properties of PCR (Section 3). Then we propose novel layers that process invariant, equivariant, and different types of bi-equivariant features, which extend standard equivariant layers by fusing information between the point clouds (Section 4).
3. **State-of-the-art in Global PCR**: We propose a novel, scalable bi-equivariant pipeline for point cloud registration. Our method ensures consistent registration results, regardless of the pose of the input point clouds, and achieves state-of-the-art registration accuracy in the robust setting, especially in low-overlap datasets.

2. Related Work

Point cloud registration (PCR) is a fundamental problem with extensive literature. Here we focus on related work on rigid geometric PCR where only depth input is provided.

Classic Methods; ICP and Global Registration. Extensive surveys (Pomerleau et al., 2015; Bellekens et al., 2015; Li et al., 2021) categorize and benchmark classical algorithms or main building blocks of those e.g., the local feature extraction backbone Guo et al. (2015) or the robust estimators Babin et al. (2018). Stemming from the Iterative Closest Point (ICP) algorithm Chen and Medioni (1991); Besl and McKay (1992), a number of variants have been proposed Pomerleau et al. (2015). The non-convexity of PCR with unknown correspondences makes ICP susceptible to local optima and thus a relatively accurate initial registration is required. This initiated the problem of *Global* PCR where PCR is treated as a global optimization problem (Li and Hartley, 2007; Yang et al., 2013) which is notoriously difficult for scene-level scans.

Correspondence-Based PCR: Correspondence-based methods utilize the local descriptors in order to match points or surfaces between the points clouds before estimating the transformation. The pioneering work of 3DMatch Zeng et al. (2017) was followed by many works that learn to match the learned keypoints (Yew and Lee, 2018; Choy et al., 2019; Sarode et al., 2019; Deng et al., 2018b; Gojcic et al., 2019; Bai et al., 2020; Wang et al., 2022; Li et al., 2020; Huang et al., 2021). More recently, keypoint-free deep learning methods have been introduced that perform matching in a coarse-to-fine fashion Yu et al. (2021); Min et al. (2021); Yang et al. (2022); Li and Harada (2022); Qi et al. (2016).

Equivariant Registration: As a step towards *global PCR*, *equivariant deep learning* can be utilized. Currently, most of the deep learning registration pipelines are not equivariant to the point cloud poses thus requiring a great amount of data augmentations Qin et al. (2022) while still behaving inconsistently during inference (Fig. 3). In this category, PPFNet Deng et al. (2018b,a) is a keypoint-based method that introduces hand-designed rotation-invariant point features as local descriptors. YOHO Wang et al. (2022) utilizes a feature extractor equivariant to the icosahedral group while SpinNet Ao et al. (2021) uses a cylindrical convolution to extract planar equivariant features. GeoTransformer Qin et al. (2022) takes a step forward by encoding pose invariant features in the superpoint transformer. However, the method is not end-to-end rotation-equivariant as we show next. Powerful rotation equivariant networks that operate on point clouds have been proposed Chen et al. (2021); Deng et al. (2021); Wu et al. (2023). They have been successfully utilized in 3D Shape Reconstruction Chatzipantazis et al. (2023); Chen et al. (2022), Segmentation Deng et al. (2023), Protein-Docking Ganea et al. (2021), Robotic Manipulation Ryu et al. (2023, 2024),Huang et al. (2024) etc. Building on the success of equivariant deep learning we propose a bi-equivariant detector-free, transformer-based PCR pipeline.

3. Problem Formulation and Characterization of Equivariant Properties

Consider a *reference* and a *source* observer, each with distinct coordinate frames r and s respectively, sampling points in their respective frames $X^r = \{x_i \in \mathbb{R}^3 | i = 1, \ldots, N\}$, $Y^s = \{y_j \in \mathbb{R}^3 | j = 1, \ldots, M\}$. Let SE(3) denote the group of roto-translations and SO(3) its subgroup of rotations. The objective of PCR (under the assumption of unique alignment) is to find the rigid transformation $\mathcal{T}_s^r \in \mathrm{SE}(3)$ that aligns the coordinate frame s to r using only the sampled points X^r, Y^s. Once the relative rotation and translation parameters $R_s^r \in SO(3), T_s^r \in \mathbb{R}^3$ that constitute $\mathcal{T}_s^r$, are estimated we can transform Y^s to the reference frame r and get $Y^r := \mathcal{T}_s^r Y^s := R_s^r Y^s + T_s^r = \{R_s^r y + T_s^r \in \mathbb{R}^3 | y \in Y^s\}$. This transformation allows the merging of the two observations with the union $X^r \cup Y^r$.

To solve this problem we assume that there exists an overlapping area of the surface sampled by both observers. Specifically, we assume that there exists a subset $X_{\mathrm{o}} \subseteq X^r$ such

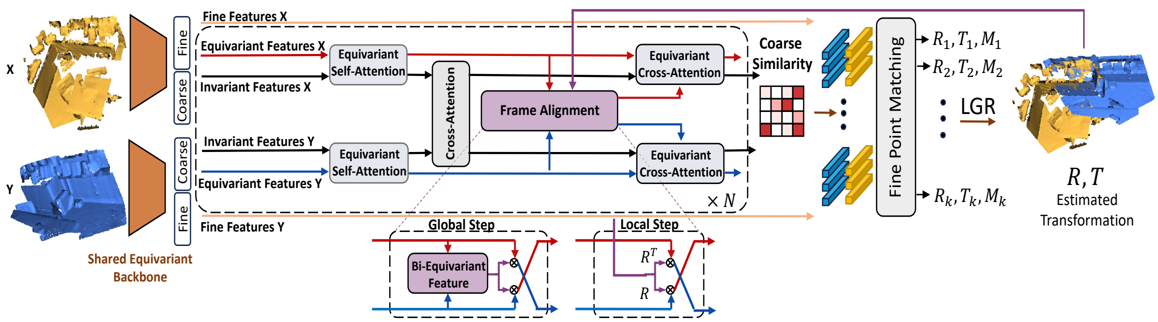

Figure 2: BiEquiFormer is an attention-based bi-equivariant pipeline for global PCR. First, equivariant intra-point self-attention and inter-point cross-attention layers update the scalar and vector features on the points. Then a bi-equivariant feature is used to align the input vectors to the same frame before applying equivariant cross-attention. The output invariant coarse features are used to extract a set of candidate coarse matches from which candidate transformations are extracted. Using these candidates the final transformation estimation is computed via a local-to-global registration scheme.

that for every point $x_m \in X_o$ there exists a corresponding $y_m \in Y^r := R^r_s Y^s + T^r_s$ such that $\|x_m - y_m\| \leq \epsilon$ for a small ϵ. We refer to the points $x_i \in X_o$ and their corresponding points $y_i \in Y^s$ as point matches. The goal is first to estimate these point matches. Given a set of such matching pairs $C = \{(x_i, y_i) | x_i \in X^r, y_i \in Y^s\}$, PCR estimates the relative transformation by solving the Procrustes optimization problem $\min_{(R,T)\in SE(3)} \sum_{(x_i,y_i)\in C} \|Ry_i + T - x_i\|_2^2$.

Characterization of Equivariant Properties of PCR: To describe the geometric properties of the problem we need the notion of *equivariance* first. Given a group G acting on two sets S_i, S_o via the (left) actions $*, \tilde{*} : G \times S \to S$ (in our cases those sets will either be (sets of) vector spaces or a sub-group of G where the action will be properly defined) a map $f : S_i \to S_o$ is *equivariant* w.r.t. the group actions if for all $g \in G, s \in S_i$: $f(g * s) = g \tilde{*} f(s)$. For clarity, we suppress $*, \tilde{*}$ and write gs for the group action of G on S.

SE(3) Bi-Equivariance: Formally, we define a function $f : S_i \to S_o$ to be SE(3)-*bi-equivariant* if it is equivariant w.r.t. the ***joint*** group action of the direct product group $\mathrm{SE}(3) \times \mathrm{SE}(3)$ defined as $s \mapsto g_1 * s \cdot g_2^{-1}$, where $*, \cdot$ are left and right group actions respectively that are jointly associative i.e. $g_1 * (s \cdot g_2^{-1}) = (g_1 * s) \cdot g_2^{-1}$ [1]. We prove in Appendix Proposition 5 that this **joint** action is a valid **left** action of the direct product group (whenever the actions $*, \cdot$ are well-defined). Depending on whether s belongs to the domain S_i or the co-domain S_o of f we define three cases . For all $(g_1, g_2) \in \mathrm{SE}(3) \times \mathrm{SE}(3)$:

Input bi-equivariance: $f : S_1 \to S_2 \times S_3$, with $f(g_1 s_1 g_2^{-1}) = (g_1 s_2, g_2 s_3), \forall s_1 \in S_1$.
Input/Output bi-equivariance: $f : S_1 \to S_2$ with $f(g_1 s_1 g_2^{-1}) = g_1 f(s_1) g_2^{-1}, \forall s_1 \in S_1$.
Output bi-equivariance, $f : S_1 \times S_2 \to S_3$, with $f(g_1 s_1, g_2 s_2) = g_1 f(s_1, s_2) g_2^{-1}$

To analyze the properties of the PCR problem we will assume that the Procrustes optimization problem has a unique solution (a sufficient condition for that is that the set C of matches includes 3 non-coplanar vectors). Given the overlap and optimality conditions

1. In our case all actions are implemented using matrix multiplications which are both left and right associative; we omit the $*, \cdot$ to make notation more compact

discussed above, and under the assumption of unique ground-truth registration $\mathcal{T}_s^r$, PCR can be defined as a map $\bigcup_{N>0} \mathbb{R}^{3\times N} \times \bigcup_{M>0} \mathbb{R}^{3\times M} \to \mathrm{SE}(3)$ with $PCR(X^r, Y^s) = \mathcal{T}_s^r$. We can rigorously prove the following propositions using our definitions (see Appendix 7.4).

Proposition 1 *PCR is* ***output*** *SE(3)-bi-equivariant. i.e. for all* $(\mathcal{T}_1, \mathcal{T}_2) \in SE(3)\times SE(3)$: $PCR(\mathcal{T}_1 X^r, \mathcal{T}_2 Y^s) = \mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1}$.

Proposition 2 *(Reference-Source Interchangeability) PCR is equivariant to the ordering of the arguments. I.e.* $C_2 = \{e, \mathfrak{f}\}$ *is the group of flips with* e *the identity and* $\mathfrak{f}$ *acting as:* $\mathfrak{f}(X^r, Y^s) = (Y^s, X^r)$ *then:* $PCR(\mathfrak{f}(X^r, Y^s)) = (\mathcal{T}_s^r)^{-1}$.

Proposition 3 *(Permutation Equivariance) PCR is invariant to the ordering of the points. I.e. if* S_N *is the group of permutations of* N *points:* $PCR(S_N X^r, S_M Y^s) = \mathcal{T}_s^r$.

4. Method

4.1. Building Bi-equivariant feature maps

While the literature is abundant with methods that build SE(3)-equivariant representations there is a lack in the design of compact and expressive bi-equivariant feature maps as described in the previous section. This is particularly important in our problem since vanilla equivariant features do not fuse the information of both point clouds thus they create impoverished representations for matching. While the general theory from Cohen et al. (2019) can be adapted to find convolutional layers, such layers have a huge memory overhead and do not scale to scene-level scans. Closer to our work, both Ganea et al. (2021) and Qin et al. (2022) parametrize only the invariant channels when they fuse the features of the point clouds via cross-attention. However, useful vector features that can be learned on the points, such as the normals of the surface, cannot be represented this way. In this work we allow for the fusion of such vector features between the two point-clouds by designing bi-equivariant layers that respect the properties described in Section 3. We present the detailed form of these layers in Section 4.2, while in Appendix 7.3 we show how their elementary operations satisfy the bi-equivariant property.

In order to mitigate this memory overhead we propose a more structured design of a bi-equivariant network as a composition of input, output and input-output bi-equivariant layers. Given our definitions above, it is straightforward to prove that we can construct an SO(3)-bi-equivariant map via the composition $(\boldsymbol{iBEq}\circ(\circ_K \boldsymbol{ioBEq}^K)\circ\boldsymbol{oBEq})(\boldsymbol{Eq}(X), \boldsymbol{Eq}(Y))$ where $\boldsymbol{iBEq}$, $\boldsymbol{oBEq}$, $\boldsymbol{ioBEq}^K$ are SO(3) input, output, and input-output bi-equivariant maps respectively and $\boldsymbol{Eq}$ is an SO(3) equivariant map.

Architecture Overview: We follow a coarse-to-fine approach similar to Qin et al. (2022). The coarse superpoint matching stage estimates candidate pairs of matching point cloud patches (superpoints). Given these, the fine point matching stage estimates R, T for the neighborhood of each candidate pair. Lastly, a local-to-global registration scheme (Appendix 7.7), is used to evaluate each candidate transformation and select the highest-scoring one. Additionally, we propose that after the first estimated transformation (Global Step) an optional Local Refinement Step can be used, using only equivariant layers. To ensure bi-equivariance all parts of the pipeline must respect the constraint. For the initial feature extractor we adapt the VNN Deng et al. (2021) architecture, as described in Appendix 7.2.

4.2. Invariant and Equivariant Attention Layers

Intra-Point Self Attention: Assume we are given a point cloud X along with its per-point equivariant and invariant features $f_s(x_i)$, $f_v(x_i)$. We propose to extend the invariant attention layer proposed in Qin et al. (2022), so that it can process both invariant and equivariant features. Specifically, we define the intra-point self-attention layers as follows:

$$\alpha_s^{\text{intra}}(x_i, f_s, f_v) = \sum_{x_j \in X} s_{ij} W_v f_s(x_j), \quad \alpha_v^{\text{intra}}(x_i, f_s, f_v) = \sum_{x_j \in X} s_{ij} \text{VN}_V(f_v(x_j))$$

where VN_V is a learned Vector Neurons linear layer and $s_{ij} = \exp(e_{ij}) / \sum_{x'_j \in X} \exp(e_{ij'})$ where e_{ij} is the attention score matrix defined as:

$$e_{ij} = (f_s(x_i)W_Q)\,(f_s(x_j)W_K + r_{ij}W_R)^T + w_q f_v(x_i)^T f_v(x_j) w_k^T$$

with r_{ij} being the invariant relative geometric embedding between x_i, x_j introduced in Qin et al. (2022), W_Q, W_K being learned weight matrices and w_q, w_k being learned weight vectors. In Appendix 7.5 we prove the invariance of α_s^{intra} and equivariance of α_v^{intra}.

Equivariant Cross-Attention Layer: Applying a mechanism similar to the intra-point self-attention for the case of inter-point cross-attention is not trivial when we want to use the equivariant features. That is because the two point clouds and their features can rotate independently, and thus an alignment is required before combining them. We propose to do such an alignment by using a bi-equivariant feature extracted from a point pair that consists of a point transforming according to frame r and a point transforming according to frame s. First, to define the point pair we assume a soft assignment $S_{XY} = \{s_{ij} \in [0,1] | \sum_{j=1}^{|Y|} s_{ij} = 1, 0 < i \leq |X|\}$ between the point clouds X and Y e.g. coming from the attention scores s_{ij} of a cross-attention layer that uses only the invariant features of the point clouds. Given S_{XY} we compute for all $x_i \in X$ the pairs (x_i, y_{pi}) where:

$$y_{pi} = \sum_{j \in |Y|} s_{ij} y_j \quad f_v(y_{pi}) = \sum\nolimits_{j \in |Y|} s_{ij} f_v(y_j) \quad f_s(y_{pi}) = \sum\nolimits_{j \in |Y|} s_{ij} f_s(y_j)$$

Then we can define the alignment layer a that aligns the equivariant features $f_v(y_{pi})$ so that they rotate according to a rotation of frame r. Specifically we define the alignment layer:

$$a(f_v(x_i), f_v(y_{pi})) = b(f_v(x_i), f_v(y_{pi})) f_v(y_{pi}) \tag{1}$$

where $b : \mathbb{R}^{3 \times C} \times \mathbb{R}^{3 \times C} \to \mathbb{R}^{3 \times 3 \times C}$ is an **output bi-equivariant** function that takes the channel-wise tensor product $f_v(x_i) \otimes f_v(y_{pi})$ and pass it through an input-output bi-equivariant nonlinearity ϕ:

$$b(f_v(x_i), f_v(y_{pi})) = \phi(f_v(x_i) \otimes f_v(y_{pi})), \quad \phi(F) = \text{LayerN}\left(\|F\|\right) \frac{F}{\|F\|}$$

with LayerN being the LayerNorm Ba et al. (2016) and $\|.\| : \mathbb{R}^{3 \times 3 \times C} \to \mathbb{R}^C$ the Frobenius norm for each 3×3 matrix. Given the set of pairs (x_i, y_{pi}) we can define the equivariant cross-attention layer where the query features are the features of points in $x_i \in X$, and the key, value features are the features of points y_{pi} after they have been properly aligned to frame r. In more detail we define the score attention matrix e_{XY}^{pair} as:

$$e_{XY}^{\text{pair}}(ij) = (f_s(x_i)W_Q)\,(f_s(y_{pj})W_K)^T + w_q f_v(x_i)^T a(f_v(x_i), f_v(y_{pj})) w_k^T$$

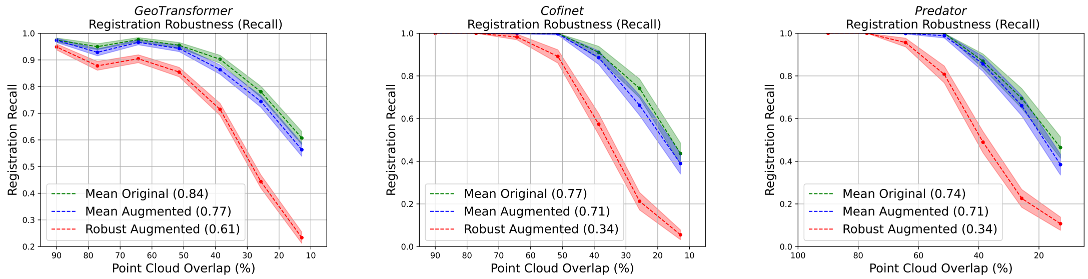

Figure 3: Registration Recall for GeoTransformer Qin et al. (2022), Cofinet Yu et al. (2021) and Predator Huang et al. (2021) on different overlap ranges of the total 3DMatch Zeng et al. (2017). The green lines (mean original) show the mean per overlap range for the original dataset. The blue lines (mean augmented) show the mean per overlap range of an augmented dataset in which each point cloud has been uniformly roto-translated. The red line (robust augmented) shows the mean per overlap range of the minimum across the augmented dataset. The total mean across all pairs in the dataset for each case is also shown in the plot.

Then assuming input invariant/equivariant features F_X,F_Y, we define the pair attention as:

$$\alpha_s^{\text{pair}}(x_i, F_X, F_Y) = \sum_{x_j \in X} s_{XY}^{\text{pair}}(ij) W_v f_s(y_{pj}),$$
$$\alpha_v^{\text{pair}}(x_i, F_X, F_Y) = \sum_{x_j \in X} s_{XY}^{\text{pair}}(ij) \left(\mathrm{VN}_V(a(f_v(x_j), f_v(y_{pj})))\right)$$

with $s_{XY}^{\text{pair}}(ij)$ being the softmax of the attention scores $e_{XY}^{\text{pair}}(ij)$. In Appendix 7.5 we prove that α_s^{pair} is invariant to the roto-translation of both point clouds X, Y. α_v^{pair} is equivariant to the roto-translation of X and invariant to the roto-translation of Y. Similarly we can define the symmetric layers for pairs of the form (y_i, x_{pi}).

4.3. BiEquiformer Architecture Stages

Coarse point correspondence: For the estimation of the superpoint matches we utilize the equivariant backbone presented in Section 7.2, followed by a coarse correspondence model that iteratively applies intra-point self-attention followed by inter-point cross attention. For the intra-point self-attention we are using in parallel the invariant and equivariant self-attention layers presented above. For the inter-point cross-attention we used a composition of a simple cross-attention layer only between the invariant features of the two point clouds, followed by a bi-equivariant cross-attention layer defined above. The outputs of the coarse correspondence transformer are the invariant per superpoint features for both point clouds, namely f_{cx}, f_{cy} for all $x \in X_S$, $y \in Y_S$. The extracted invariant features are then used similarly to Qin et al. (2022) to extract the candidate superpoint invariant matches.

Fine point matching: Given a candidate pair of matched superpoints $(x_{k(n)}, y_{k(n)})$ we perform fine point matching on their corresponding local neighborhoods $\mathcal{N}_{x_{k(n)}} \subseteq X_D$, $\mathcal{N}_{y_{k(n)}} \subseteq Y_D$. We define the neighborhood $\mathcal{N}_{x_{k(n)}} \subseteq X_D$ as the set of all the fine points that have $x_{k(n)}$ as their closest coarse point, and similarly for $\mathcal{N}_{y_k}$. The dense point correspondences are extracted using an optimal transport layer (Sinkhorn and Knopp, 1967) with a

Model	Canonical RR 3DM	Canonical RR 3DLM	Roto-translated Mean RR 3DM+3DLM	Robust RR 3DM+3DLM	Mean IR 3DM+3DLM	Robust IR 3DM+3DLM
FCGF Choy et al. (2019)	0.85	0.40	-	-	-	-
D3Feat Bai et al. (2020)	0.82	0.37	-	-	-	-
Predator Huang et al. (2021)	0.89	0.60	0.71	0.34	0.36	0.25
CoFiNet Yu et al. (2021)	0.89	0.68	0.71	0.34	0.38	0.27
GeoTransformer Qin et al. (2022)	0.91	0.74	0.77	0.61	0.49	0.46
LepardLi and Harada (2022)	0.92	0.65	0.64	0.60	0.37	0.30
SpinNet Ao et al. (2021)	0.89	0.60	0.72	-	0.36	-
YOHO Wang et al. (2022)	0.90	0.65	0.76	-	0.43	-
RIGA Yu et al. (2024)	0.89	0.65	0.77	0.77	0.47	0.47
BiEquiformer	0.90	0.69	0.78	0.78	0.49	0.49

Table 1: Top: Non-equivariant methods, Bottom: Equivariant methods. Canonical Registration Recall (RR) on 3DMatch (3DM) and 3DLoMatch (3DLM), Mean and Robust Registration Recall (Mean RR, Robust RR) and Inlier Ratio (Mean IR, Robust IR) on the total Rotated 3DMatch (concatenation of the 3DMatch and 3DLoMatch) for inputs augmented by uniform rotation.

cost matrix defined as $C_k = \left(F_{x_k} F_{y_k}^T\right)/\sqrt{d}$. These dense correspondences define a set of inliers M_k for each candidate pair. Here $F_{x_k} \in R^{C\times|\mathcal{N}_{x_k}|}$, $F_{y_k} \in R^{C\times|\mathcal{N}_{y_k}|}$ are matrices with columns containing scalar features for each point of the corresponding local neighborhoods. Similar to the coarse matches, in order for the optimal transport cost and consequently the assignment of the fine point matches to be invariant to rigid transformation, the features F_{x_k}, F_{y_k} should also be invariant to these transformations. In our design we achieve that by concatenating the extracted invariant features and the invariant inner products between the extracted equivariant features. After the dense point correspondences are computed the final alignment transformation is estimated using a local-to-global registration scheme proposed in Qin et al. (2022) (See Appendix 7.7).

Iterative Refinement: Given an initial estimation of the alignment transformation R_0, T_0 produced by our model, we can perform a refinement step by iteratively applying our model. Specifically after the first iteration we can use the previously estimated transformation to replace the bi-equivariant feature $b(f_v(x_i), f_v(y_{pi}))$ used in the alignment layer defined in Eq. 1. In the experimental results we perform three such refinement steps.

5. Experiments

We evaluate our method on the 3DMatch Zeng et al. (2017) and the challenging 3DLoMatch Huang et al. (2021) datasets which contain scans of indoor scenes. Following Huang et al. (2021) we evaluate on the 3DMatch test set containing scenes with an overlap above 30% and on the 3DLoMatch test set, which contains scenes with overlap from 10% to 30%. For the quantitative evaluation of our method we use similar metrics to previous works Qin et al. (2022); Huang et al. (2021) (see Appendix 7.8 for more details).

5.1. Robustness Analysis to the initial pose of the point clouds

We benchmark popular state-of-the-art methods Qin et al. (2022); Yu et al. (2021); Huang et al. (2021) on their robustness to the initial poses of the scans. We test all methods in the total 3DMatch dataset Zeng et al. (2017) by concatenating the 3DMatch and 3DLoMatch splits and test the mean performance across different overlap intervals. In Fig.3 we plot the

Registration Recall in 3 different settings. First, in the *mean original*, we show the mean performance of the methods in each overlap interval in the original dataset. The overlap of each pair is calculated as in Huang et al. (2021) in the ground truth registration. Second, in the *mean augmented*, we show the mean performance in an augmented dataset where each point cloud from each pair has been individually rotated around 9 axes uniformly selected and with 3 different angles around each axis also uniformly selected. Lastly, in the *robust augmented*, we show the robust loss i.e., the mean of the minimum performance for each overlap region across the different configurations of the same pair. We observe that there is a big drop in performance in the augmented setting both in the average (6-7%) and in the robust metrics (23-43%), which is exacerbated as the overlap becomes smaller. While GeoTransformer is more robust to initial poses than the rest of the methods it still performs erratically in different initial poses (see Figure 1). This is because although the model used for the superpoint matching is designed to be invariant to point-cloud poses, its backbone Thomas et al. (2019) is not rotation equivariant. Nevertheless, the improved robustness shows that enforcing equivariance even partially in the model can be beneficial for global PCR. These results motivate the design of our end to end bi-equivariant PCR pipeline.

5.2. Quantitative Comparison

We compare the performance of our method against recent state-of-the-art, FCGF Choy et al. (2019), D3Feat Bai et al. (2020), SpinNet Ao et al. (2021), Predator Huang et al. (2021),YOHO Wang et al. (2022), CoFiNet Yu et al. (2021), GeoTransformer Qin et al. (2022). Details on the training and evaluation metrics are presented in Appendix 7.6.2, 7.8. In Table 1 we present the Registration Recall separately for the original 3DMatch and 3DLoMatch. Then, in order to measure robustness to the initial poses of the point clouds, which is the important metric for *global* PCR, we estimate the *expected registration recall* (Mean RR) across different initial poses and the *robust registration recall* which is the average over the dataset of the minimum recall over different poses of the input, similar to the experiment in Section 5.1. We observe that our method achieves comparable results with other state-of-the-art methods in the canonical test set, being second only to GeoTransformer. Moreover, it achieves state-of-the-art performance in the expected and robust metrics. This validates the argument that our bi-equivariant design is an important step towards *global* PCR without sacrificing performance on the canonical setting. Visualizations of low-overlap registrations are provided in Appendix Fig. 4, while in Appendix 7.1 we provide additional ablations experiments on the proposed bi-equivariant layers.

6. Conclusion

In this work we proposed a novel bi-equivariant pipeline to address the task of *global* PCR. We investigated the robustness of current deep learning methods on the transformation of the poses of the input scans and observed a large performance degradation, especially in low-overlap settings. To combat that we proposed to build novel, expressive bi-equivariant layers that fuse the information of the two point clouds while extracting per-point features on them. We used those layers to build BiEquiformer a bi-equivariant attention architecture that is scalable to the large volume of points in scene-level scans. We evaluated our method on both the 3DMatch and the challenging 3DLoMatch dataset, showing that our method can achieve comparable and even superior performance to other non-equivariant and equivariant state-of-the-art methods, especially in the robust metrics.

References

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*, 2021.

Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust and efficient point cloud registration using pointnet. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Philippe Babin, Philippe Giguère, and F. Pomerleau. Analysis of robust functions for registration algorithms. *2019 International Conference on Robotics and Automation (ICRA)*, pages 1451–1457, 2018. URL https://api.semanticscholar.org/CorpusID:52912585.

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. *arXiv:2003.03164 [cs.CV]*, 2020.

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. *CVPR*, 2021.

Ben Bellekens, Vincent Spruyt, Raf Berkvens, Rudi Penne, and Maarten Weyn. A benchmark survey of rigid 3d point cloud registration algorithms. *International Journal On Advances in Intelligent Systems*, 1, 06 2015.

P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

G. Blais and M.D. Levine. Registering multiview range data to create 3d computer objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):820–824, 1995. doi: 10.1109/34.400574.

Michael Bosse and Robert Zlot. Map matching and data association for large-scale two-dimensional laser scan-based slam. *I. J. Robotic Res.*, 27:667–691, 06 2008. doi: 10.1177/0278364908091366.

Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Edgar Dobriban, and Kostas Daniilidis. Se(3)-equivariant attention networks for shape reconstruction in function space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RDy3IbvjMqT.

Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. pages 14514–14523, 2021.

Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, pages 2724–2729 vol.3, 1991. doi: 10.1109/ROBOT.1991.132043.

Yunlu Chen, Basura Fernando, Hakan Bilen, Matthias Nießner, and Efstratios Gavves. 3d equivariant graph implicit functions. *ECCV*, 2022.

Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, 2019.

Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Taco S. Cohen, Mario Geiger, and Maurice Weiler. *A general theory of equivariant CNNs on homogeneous spaces*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas Guibas. Vector neurons: a general framework for so(3)-equivariant networks. *arXiv preprint arXiv:2104.12229*, 2021.

Congyue Deng, Jiahui Lei, Bokui Shen, Kostas Daniilidis, and Leonidas J. Guibas. Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance. *ArXiv*, abs/2305.16314, 2023. URL https://api.semanticscholar.org/CorpusID:258887967.

Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. *ArXiv*, abs/1808.10322, 2018a. URL https://api.semanticscholar.org/CorpusID:52131369.

Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018b. URL https://api.semanticscholar.org/CorpusID:3703761.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Independent se (3)-equivariant models for end-to-end rigid protein docking. *arXiv preprint arXiv:2111.07786*, 2021.

Zan Gojcic, Caifa Zhou, Jan Dirk Wegner, and Wieser Andreas. The perfect match: 3d point cloud matching with smoothed densities. In *International conference on computer vision and pattern recognition (CVPR)*, 2019.

Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Kwok. A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116, 04 2015. doi: 10.1007/s11263-015-0824-y.

Haojie Huang, Owen Lewis Howell, Dian Wang, Xupeng Zhu, Robert Platt, and Robin Walters. Fourier transporter: Bi-equivariant robotic manipulation in 3d. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UulwvAU1W0.

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 (CVPR)*, pages 4267–4276, June 2021.

Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Hongdong Li and Richard Hartley. The 3d-3d registration problem revisited. pages 1 – 8, 11 2007. doi: 10.1109/ICCV.2007.4409077.

Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *European Conference on Computer Vision (ECCV)*, 2020.

Leihui Li, Riwei Wang, and Xuping Zhang. A tutorial review on point cloud registrations: Principle, classification, comparison, and technology challenges. *Mathematical Problems in Engineering*, 2021. URL https://api.semanticscholar.org/CorpusID:237702700.

Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Taewon Min, Chonghyuk Song, Eunseok Kim, and Inwook Shim. Distinctiveness oriented positional equilibrium for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5490–5498, October 2021.

Andreas Nüchter, Kai Lingemann, Joachim Hertzberg, and Hartmut Surmann. 6d slam - 3d mapping outdoor environments. *Fraunhofer IAIS*, 24, 11 2006.

G. Dias Pais, Pedro Miraldo, Srikumar Ramalingam, Jacinto C. Nascimento, Venu Madhav Govindu, and Rama Chellappa. 3dregnet: A deep neural network for 3d point registration. pages 7193–7203, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

François Pomerleau, Francis Colas, and Roland Siegwart. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends® in Robotics*, 4:1–104, 05 2015. doi: 10.1561/2300000035.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2016. URL http://arxiv.org/abs/1612.00593. cite arxiv:1612.00593.

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 (CVPR)*, pages 11143–11152, June 2022.

Hyunwoo Ryu, Hong in Lee, Jeong-Hoon Lee, and Jongeun Choi. Equivariant descriptor fields: Se(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dnjZSPGmY5O.

Hyunwoo Ryu, Jiwoo Kim, Hyunseok An, Junwoo Chang, Joohwan Seo, Taehan Kim, Yubin Kim, Chaewon Hwang, Jongeun Choi, and Roberto Horowitz. Diffusion-edfs: Bi-equivariant denoising generative modeling on se(3) for visual robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18007–18018, June 2024.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. URL https://arxiv.org/abs/1911.11763.

Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding, 2019.

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

Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6410–6419, 2019. doi: 10.1109/ICCV.2019.00651.

Haiping Wang, Yuan Liu, Zhen Dong, and Wenping Wang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1630–1641, 2022.

Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019.

Wenxuan Wu, Li Fuxin, and Qi Shan. Pointconvformer: Revenge of the point-based convolution. In *CVPR*, 2023. URL https://arxiv.org/abs/2208.02879.

Fan Yang, Lin Guo, Zhi Chen, and Wenbing Tao. One inlier is first: Towards efficient position encoding for point cloud registration. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=19MmorTQhho.

Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-icp: Solving 3d registration efficiently and globally optimally. In *2013 IEEE International Conference on Computer Vision*, pages 1457–1464, 2013. doi: 10.1109/ICCV.2013.184.

Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *ECCV*, 2018.

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, 2021.

Hao Yu, Ji Hou, Zheng Qin, Mahdi Saleh, Ivan Shugurov, Kai Wang, Benjamin Busam, and Slobodan Ilic. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3796–3812, 2024. doi: 10.1109/TPAMI.2023.3349199.

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.

7. Appendix / Supplementary Material

7.1. Ablation on the proposed layer and the iterative refinement

In this section we provide an ablation study, presented in table 2, to show the importance of the proposed bi-equivariant layers as well as the proposed equivariant iterative refinement. First, we provide a simple end-to-end equivariant alternative to GeoTransformer by replacing the non-equivariant feature extractor KPConv Thomas et al. (2019) with the equivariant VNN Deng et al. (2021). We show that when compared to this simple alternative, BiEquiFormer which uses bi-equivariant layers that fuse the information from the two point clouds demonstrates improved performance on the task. Moreover, we experimented with local refinement steps after the initial global alignment. We ran the non-equivariant ICP algorithm, heavily tuned (Point-to-Plane ICP with Robust loss Pomerleau et al. (2015)). Then we ran the equivariant iterative scheme described in Section 4.3. In this case too, our method yields better results.

Model	RR	
	3DM	3DLM
VNN+GeoTransformer	0.87	0.62
BiEquiformer + ICP	0.88	0.66
BiEquiFormer	0.90	0.69

Table 2: Ablation study on BiEquiformer. VNN+GeoTransformer replaces the non-equivariant KPConv Thomas et al. (2019) with an equivariant counterpart VNN Deng et al. (2021). BiEquiFormer+ICP utilizes the bi-equivariant layers but refines with a non-bi-equivariant ICP. BiEquiFormer uses the equivariant iterative scheme described in Section 4.3

7.2. Equivariant Feature Extraction

Previous works utilize commonly used point cloud processing architectures, such as KPConv-FPN (Thomas et al., 2019) or DGCNN (Wang et al., 2019), to extract per point features for each point-cloud individually. These features are not inherently designed to be equivariant to rigid transformations. We address this limitation by using a backbone feature extractor that outputs both invariant f_s and equivariant f_v feature vectors. Under a roto-translation R, T of the input these features transform as:

$$f_s(Rx_i + T, RX + T) = f_s(x_i, X), \qquad f_v(Rx_i + T, RX + T) = Rf_v(x_i, X) \tag{2}$$

To process such equivariant vector features we utilize the Vector Neurons layer proposed in Deng et al. (2021). This type of linear layer, denoted as VN, processes features of the form $F \in \mathbb{R}^{3\times C}$, with columns corresponding to vectors in $\mathbb{R}^3$. It is defined as $\mathrm{VN}(F) = FW_{lvn}$, and is equivariant to rotations of its input features since $\mathrm{VN}(RF) = RFW_{lvn} = R\mathrm{VN}(F)$.

Additionally, to capture the geometry of the scenes at different levels of detail we use a hierarchical architecture, similar to Chen et al. (2022), that processes and outputs invariant/equivariant vector features for different subsampled versions of the input point cloud. We denote these subsampled versions as $X_{(0)}, X_{(1)}, \ldots, X_{(n)}$, ranging from finer to coarser sampled points. We can create the different levels by running for example an equivariant adaptation of Farthest Point Sampling (FPS) where we initialize it from the point closest to the mean. The points in the first downsampling level are referred as dense points $X_D = X_{(1)}$, while the points obtained by the last level of downsampling are referred to as superpoints $X_S = X_{(n)}$.

7.3. Bi-Equivariant Layers

In this section, we present examples of simple operations that preserve the three cases of bi-equivariance, namely the input-bi-equivariance the output-bi-equivariance and the input-output-bi-equivariance:

Proposition 4

1. *If $f_1, f_2 \in \mathbb{R}^3$ are vector features i.e. they transform with the standard representation of $SO(3)$ then the tensor product $f_1, f_2 \mapsto f_1 f_2^T$ is an SO(3)* ***output-bi-equivariant*** *map.*
2. *Given a matrix $F \in \mathbb{R}^{3\times 3}$ with distinct, positive, singular values that transforms with the joint action of $SO(3) \times SO(3)$ i.e., $F \mapsto R_1 F R_2^T$ the map:*

$$F \mapsto (\{U_i \sigma(\Sigma)\}_{i=1}^4, \{V_i \sigma'(\Sigma)\}_{i=1}^4)$$

is an SO(3) ***input-bi-equivariant map****, where $\{(U_i, \Sigma, V_i)\}_{i=1}^4$ are the 4 possible SVD decompositions of F counting signs with $U_i, V_i \in SO(3)$ if $\det(F) > 0$ and $U_i \in O(3) - SO(3), V_i \in SO(3)$ if $\det(F) < 0$ and σ, σ' are point-wise non-linearities on the singular values. The $SO(3)$ matrices are formed as $[u_1, u_2, u_1 \times u_2]$ and the $O(3) - SO(3)$ as $[u_1, u_2, -u_1 \times u_2]$ where u_1, u_2 are the first two columns of U (and similar for V).*

3. *Given the same matrix $F \neq 0$ as above, the map $F \mapsto \sigma(\|F\|)\frac{F}{\|F\|}$ is $\mathrm{SO}(3)$* ***input-output bi-equivariant****, where $\|\cdot\|$ is a matrix norm e.g. operator, Frobenius, trace norm etc.*

We provide a detailed proof of the above proposition along with the proofs for propositions 1, 2, 3, 5 in the next Section 7.4.

By composing the above operations we can design bi-equivariant layers that allow for fusion of information between equivariant features that transform according to independent frames. This capability of bi-equivariant layers makes them ideal for the problem of global PCR, contrary to the traditional equivariant layers that can only process features expressed in the same frame. Specifically, as described in Section 4.2, BiEquiformer uses both the tensor product and the bi-equivariant map $F \mapsto \sigma(\|F\|)\frac{F}{\|F\|}$ to define the inter-point cross attention layers.

7.4. Proofs of Propositions

Before beginning with the proofs of the propositions we need to prove a subtle but important point that the joint action is indeed a valid group action of the direct product group.

Proposition 5 *If the groups G_1, G_2 act on the set S via $*, \cdot$ from the right and the left respectively and these actions are jointly associative i.e. $(g_1 * s).g_2 = g_1 * (s.g_2)$, for all*

$g_1 \in G_1, g_2 \in G_2, s \in S$ then the map defined as:

$$\begin{aligned}
&(G_1 \times G_2) \times S \to S \\
&((g_1, g_2), s) \mapsto g_1 * (s \cdot g_2^{-1})
\end{aligned}$$

is a group action of the direct product group $G_1 \times G_2$.

Proof We write the map as $(g_1, g_2)s := g_1 * (s \cdot g_2^{-1})$ for compactness. If e_1, e_2 are the identity elements of G_1, G_2 then (e_1, e_2) is the identity element of $G_1 \times G_2$. Also consider $(g_1, g_2), (h_1, h_2) \in G_1 \times G_2$ Then,

$$\begin{aligned}
1.&(e_1, e_2)s = e_1 * (s \cdot e_2^{-1}) = e_1 * (s \cdot e_2) = e_1 * s = s \\
2.&(g_1, g_2)(h_1, h_2)s = (g_1, g_2)(h_1 * (s \cdot h_2^{-1})) = g_1 * ((h_1 * (s \cdot h_2^{-1})) \cdot g_2^{-1}) \overset{\text{joint assoc.}}{=} \\
&= g_1 * (h_1 * ((s \cdot h_2^{-1}) \cdot g_2^{-1})) = g_1 * (h_1 * ((s \cdot h_2^{-1} g_2^{-1}))) = (g_1 h_1) * (s \cdot (h_2^{-1} g_2^{-1})) \\
&= (g_1 h_1, g_2 h_2)s
\end{aligned}$$

■

Due to joint associativity we can drop the parentheses and write $(g_1, g_2)s := g_1 * s \cdot g_2^{-1}$. We did not do that in the proof to make explicit when the joint associativity was used.

Proof [Proof of Proposition 1] Given the formulation in Section 3 we start by denoting the input point clouds X^r, Y^s and their relative rigid transformation $\mathcal{T}_s^r = \begin{bmatrix} R_s^r & T_s^r \\ 0 & 1 \end{bmatrix}$. Also let $C = \{(x_i, y_i) | x_i \in X^r, y_i \in Y^s\}$ denote the point matches. Now, if the input point clouds transform with $\mathcal{T}_1, \mathcal{T}_2 \in \mathrm{SE}(3)$ as: $\mathcal{T}_1 X^r = R_1 X^r + T_1$, $\mathcal{T}_2 Y^s = R_2 Y^s + T_2$ then we need to prove the following for the transformation $\mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1} \in \mathrm{SE}(3)$:

- Invariant point matching: The points $\mathcal{T}_1 x_i = R_1 x_i + T_1 \in \mathcal{T}_1 X^r, \mathcal{T}_2 y_i = R_2 y_i + T_2 \in \mathcal{T}_2 Y^s$ are also point matches for $\mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1}$ (which can also be computed from the first problem formulation) since in the new alignment we have: $\mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1}(\mathcal{T}_2 y_i) = \mathcal{T}_1 \mathcal{T}_s^r y_i$ and

$$\begin{aligned}
\|\mathcal{T}_1 x_i - \mathcal{T}_1 \mathcal{T}_s^r y_i\|_2 &= \|(R_1 x_i + T_1) - (R_1(\mathcal{T}_s^r y_i) + T_1)\|_2 \\
&= \|R_1(x_i - \mathcal{T}_s^r y_i)\|_2 = \|x_i - \mathcal{T}_s^r y_i\|_2 \leq \epsilon
\end{aligned}$$

  since $(x_i, y_i) \in C$.

- Optimal Procrustes: For the initial problem we know that the objective function $L_1(\mathcal{T}) = \sum_{(x_i, y_i) \in C} \|\mathcal{T} y_i - x_i\|_2^2$ satisfies: $L_1(\mathcal{T}_s^r) := L_1^* \leq L_1(\mathcal{T})$ for all $\mathcal{T} \in \mathrm{SE}(3)$. Now we look at the objective of the new problem (for which we proved invariant matches) $L_2(\mathcal{T}) = \sum_{(x_i, y_i) \in C} \|\mathcal{T} \mathcal{T}_2 y_i - \mathcal{T}_1 x_i\|_2^2$. If we substitute $\mathcal{T} = \mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1}$ we get:

$$L_2(\mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1}) = \sum_{(x_i, y_i) \in C} \|\mathcal{T}_1 \mathcal{T}_s^r \mathcal{T}_2^{-1} \mathcal{T}_2 y_i - \mathcal{T}_1 x_i\|_2^2 = \sum_{(x_i, y_i) \in C} \|\mathcal{T}_s^r y_i - x_i\|_2^2 = L_1(\mathcal{T}_s^r) = L_1^*$$

we proved that the optimal of the second problem is upper bounded by the first. We will also show the opposite. In particular, if we substitute $\mathcal{T} = \mathcal{T}_1^{-1}\mathcal{T}\mathcal{T}_2$ in L_1 for any $\mathcal{T} \in \mathrm{SE}(3)$ we get:

$$\begin{aligned}
L_1(\mathcal{T}_1^{-1}\mathcal{T}\mathcal{T}_2) &= \sum_{(x_i,y_i)\in C} \|\mathcal{T}_1^{-1}\mathcal{T}\mathcal{T}_2 y_i - x_i\|_2^2 \\
&= \sum_{(x_i,y_i)\in C} \|R_1^T(\mathcal{T}\mathcal{T}_2 y_i) - R_1^T T_1 - x_i\|_2^2 \\
&= \sum_{(x_i,y_i)\in C} \|\mathcal{T}\mathcal{T}_2 y_i - R_1(R_1^T T_1 + x_i)\|_2^2 \\
&= \sum_{(x_i,y_i)\in C} \|\mathcal{T}\mathcal{T}_2 y_i - \mathcal{T}_1 x_i\|_2^2 = L_2(\mathcal{T})
\end{aligned}$$

■

Proof [Proof of proposition 2] First, we can again prove invariant matching. The flip is a unitary operation so it does not change the distances between the matched points. In other words since $\|x_m - y_m\|_2 = \|y_m - x_m\|_2$ the set C of point matches consists of the same points (reversed). Again looking at the two objectives we can prove Procrustes optimality as for $\mathcal{T} \in \mathrm{SE}(3)$ it holds $\mathcal{T}^{-1} \in \mathrm{SE}(3)$:

$$\begin{aligned}
L_1(\mathcal{T}^{-1}) &= \sum_{(x_i,y_i)\in C} \|\mathcal{T}^{-1} y_i - x_i\|_2^2 \\
&= \sum_{(x_i,y_i)\in C} \|R^T y_i - R^T T - x_i\|_2^2 \\
&= \sum_{(x_i,y_i)\in C} \|y_i - T - R x_i\|_2^2 = \sum_{(x_i,y_i)\in C} \|\mathcal{T} x_i - y_i\|_2^2 = L_2(\mathcal{T})
\end{aligned}$$

Thus, the optimal values of the two problems are again the same and since $\mathcal{T}_s^r$ is optimal for L_1 then $(\mathcal{T}_s^r)^{-1}$ is optimal for L_2. Lastly, this is indeed an action of the flips since $f^2 = e$ and $((\mathcal{T}_s^r)^{-1})^{-1} = \mathcal{T}_s^r$ ■

Proof [proof of Proposition 3] Since the permutations is a unitary transformation the distance again as above do not change and the matching is again invariant (this time the set has exactly the same points in some order). Since the sum is order-invariant the value of the objective is also the same so the problem is invariant to point permutations. ■

Proof [proof of proposition 4]

1. Since $f_1 \mapsto R_1 f_1, f_2 \mapsto R_2 f_2$ the tensor product $f_1 f_2^T \mapsto (R_1 f_1)(R_2 f_2)^T = R_1(f_1 f_2^T)R_2^T$. Thus, the map is output bi-equivariant.
2. Since all singular values are distinct and positive, we can sort them in $\Sigma = diag\{\sigma_1, \sigma_2, \sigma_3\}$ in which case it is known that the SVD of $F = U\Sigma V^T$ is unique up to a simultaneous sign flip of the columns of U, V i.e., there are 8 choices for $U = [\pm u_1 \;\; \pm u_2 \pm u_3]$

and the corresponding for V. However, if $\det(F) > 0$ then we can select both $U, V \in SO(3)$ i.e. $u_3 = u_1 \times u_2$ and $v_3 = v_1 \times v_2$ and if $\det(F) < 0$ we can select $U \in O(3) - SO(3), V \in SO(3)$ i.e. $u_3 = -u_1 \times u_2, v_3 = v_1 \times v_2$. Since all singular values are positive the determinant cannot be zero.

That leaves 4 choices i.e. if (u_1, v_1) and (u_2, v_2) are the first and second columns of U, V then $\pm(u_1, v_1)$, $\pm(u_2, v_2)$ are the rest of the choices for the first and second column of U, V which create the valid SVD solutions.

Now, $R_1 F R_2^T = R_1(U\Sigma V^T)R_2^T = (R_1U)\Sigma(R_2V)^T$ and thus (R_1U, Σ, R_2V) is an SVD of $R_1FR_2^T$ since the composition of rotation matrix with a unitary matrix is a unitary matrix. Moreover, $\det(R_1FR_2^T) = \det(R_1)\det(F)\det(R_2) = \det(F)$ so if $\det(F) > 0$ then $R_1U, R_2V \in SO(3)$ and if $\det(F) < 0$ then $R_1U \in O(3) - SO(3), R_2V \in SO(3)$ as is the case for U, V.

So if the set $\{(U_1, V_1), (U_2, V_2), (U_3, V_3), (U_4, V_4)\}$ is the set of valid U, V in the SVD for F then for $R_1FR_2^T$ the corresponding set is: $\{(R_1U_1, R_2V_1), (R_1U_2, R_2V_2), (R_1U_3, R_2V_3), (R_1U_4, R_2V_4)\}$. Also, Σ is invariant. Thus we can use any point-wise non-linearity on Σ since this is also invariant. And if we define the action $\cdot$ on the set of 4 matrices as $R \cdot \{U_1, U_2, U_3, U_4\} = \{RU_1, RU_2, RU_3, RU_4\}$ then the map:

$$F \mapsto (\{U_i\sigma(\Sigma)\}_{i=1}^4, \{V_i\sigma'(\Sigma)\}_{i=1}^4)$$

satisfies:

$$R_1FR_2^T \mapsto (R_1 \cdot \{U_i\sigma(\Sigma)\}_{i=1}^4, R_2 \cdot \{V_i\sigma'(\Sigma)\}_{i=1}^4).$$

Thus the map is input bi-equivariant.

3. Since $\|R_1FR_2^T\| = \|F\|$ we get $R_1FR_2^T \mapsto \sigma(\|R_1FR_2^T\|)\frac{R_1FR_2^T}{\|R_1FR_2^T\|} = R_1\sigma(\|F\|)\frac{F}{\|F\|}R_2^T$. Thus the map is input-output bi-equivariant.

■

7.5. Equivariant and Bi-Equivariant Properties of Attention layers

In this section we provide a more detailed analysis of the properties of the attention layers defined in Section 4.2.

Intra point self-attention layer: Since this layer processes points coming from the same point cloud and thus features that transform according to the same frame we require it to preserve the standard equivariant property. Specifically, we can show the following:

Proposition 6 *α_s^{intra} is invariant and α_v^{intra} is equivariant to the roto-translation of the input point cloud:*

$$\alpha_s^{\text{intra}}(Rx_i + T, f_s, Rf_v) = \alpha_s^{\text{intra}}(x_i, f_s, f_v)$$
$$\alpha_v^{\text{intra}}(Rx_i + T, f_s, Rf_v) = R\alpha_v^{\text{intra}}(x_i, f_s, f_v)$$

Proof [Proof sketch of proposition] It is easy to show that e_{ij} is invariant to transformations of all the inputs of a_s^{intra} since the first term uses only the invariant f_s features and the invariant r_{ij} geometric embedding introduced in Qin et al. (2022). In the second term a transformation by R results in:

$$w_q(Rf_v(x_i))^T(Rf_v(x_j))w_k = w_q f_v(x_i)^T R^T R f_v(x_j) w_k = w_q f_v(x_i)^T f_v(x_j) w_k$$

which is also invariant. As a result $\alpha_s^{\text{intra}}(x_i, f_s, f_v)$ is invariant since it only depends on e_{ij} and f_v and since the VN layer is equivariant to the rotations:

$$\begin{aligned}
\alpha_v^{\text{intra}}(Rx_i + T, f_s, Rf_v) = &= \sum_{x_j \in X} \frac{\exp(e_{ij})}{\sum_{x'_j \in X} \exp(e_{ij'})} \mathrm{VN}_V(Rf_v(x_j)) \\
&= \sum_{x_j \in X} \frac{\exp(e_{ij})}{\sum_{x'_j \in X} \exp(e_{ij'})} R\mathrm{VN}_V(f_v(x_j)) \\
&= R\alpha_v^{\text{intra}}(x_i, f_s, f_v)
\end{aligned}$$

■

Inter point cross-attention: In this case the inter point cross-attention layer is required to process and fuse information between points that transform according to different frames. Thus it is required to preserve the bi-equivariant properties presented in Section 3. Starting from the alignment layer of Equation 1 we can show the following:

Proposition 7 *The alignment layer is equivariant to the rotations of its first input and invariant to the rotations of its second input:* $a(R_x f_v(x_i), R_y f_v(y_{pi})) = R_x a(f_v(x_i), f_v(y_{pi}))$

Proof [Proof Sketch] Here following Proposition 4 we use the fact that the the Frobenius norm is invariant to the rotation and as a result for the nonlinearity we have that:

$$\begin{aligned}
\phi(R_x F R_y^T) &= \mathrm{LayerN}\left(\|R_x F R_y^T\|\right) \frac{R_x F R_y^T}{\|R_x F R_y^T\|} \\
&= \mathrm{LayerN}\left(\|F\|\right) \frac{R_x F R_y^T}{\|F\|} \\
&= R_x \phi(F) R_y^T
\end{aligned}$$

Then using the fact that the tensor product is bi-equivariant it is easy to show that b is **output bi-equivariant**:

$$\begin{aligned}
b(R_x f_v(x_i), R_y f_v(y_{pi})) &= \phi(R_x f_v(x_i) \otimes R_y f_v(y_{pi})) \\
&= \phi(R_x(f_v(x_i) \otimes f_v(y_{pi})) R_Y^T) \\
&= R_x \phi((f_v(x_i) \otimes f_v(y_{pi}))) R_Y^T
\end{aligned}$$

and then

$$\begin{aligned}
a(R_x f_v(x_i), R_y f_v(y_{pi})) &= b(R_x f_v(x_i), R_y f_v(y_{pi})) R_y f_v(y_{pi}) \\
&= R_x b(f_v(x_i), f_v(y_{pi})) f_v(y_{pi}) \\
&= R_x a(f_v(x_i), f_v(y_{pi}))
\end{aligned}$$

■

Finally, for the overall inter-point cross-attention layer we can show that:

Proposition 8 *α_s^{pair} is invariant to the roto-translation of both point clouds X, Y. α_v^{pair} is equivariant to the roto-translation of X and invariant to the roto-translation of Y. Specifically given $X' = R_x X + T_x$ and $Y' = R_y Y + T_Y$:*

$$\alpha_s^{\text{pair}}(R_x x_i + T_x, F_{X'}, F_{Y'}) = \alpha_s^{\text{pair}}(x_i, F_X, F_Y)$$
$$\alpha_s^{\text{pair}}(R_x x_i + T_x, F_{X'}, F_{Y'}) = R_X \alpha_s^{\text{pair}}(x_i, F_X, F_Y)$$

Proof [Proof sketch] Here the layer is similar with the one in proposition 4.1 with different second input being $a(f_v(x_i), f_v(y_{pi}))$ that is equivariant to the transformation of frame X. So we can show the equivariance using the same arguments as proposition 4.1 ■

7.6. Implementation Details

7.6.1. Input pre-processing

For the initial feature extraction, described in Section 7.2, we use four different subsampled versions of the input point cloud, denoted as $X^{(0)}, X^{(1)}, X^{(2)}, X^{(3)}$. Each point cloud is sampled using grid sampling where, for the i^{th} subsampled version $X^{(i)}$, the voxel size is set to $0.025 * 2^i$.
During training, both the source and the reference point clouds are augmented with Gaussian noise with standard deviation of 0.005. Additionally, for each point cloud, we limit the total amount of points to 5000. If the input point clouds exceed this limit, we randomly sample 5000 points from each one of them. We observed that enforcing this limit during training has a minimum effect on the performance during testing, even when we test on larger point clouds.

7.6.2. Model Architecture and Training

We implemented and evaluated BiEquiFormer in PyTorch Paszke et al. (2019) on an I9 Intel CPU, 64GB RAM and an NVIDIA RTX3090 GPU.

- **Feature extraction**: Our feature extraction network consists of consecutive "hybrid" layers, similar to the ones proposed in Chen et al. (2022), that simultaneously process both scalar invariant features and equivariant vector features by utilizing Vector Neurons layers (Deng et al., 2021). In each layer, all points aggregate features from their k nearest neighbors, where we set $k = 20$. We perform three aggregation steps for each subsampled version of the point cloud. Similar to KPConv-FPN (Thomas et al., 2019), we process the different subsampled versions from finer to coarser, where the coarser points have as input features an aggregation of the extracted features of their closest finer points.

- **Coarse point correspondence**: The coarse point correspondence model consists of three consecutive blocks of an intra-point self-attention layer described in Section **??**, followed by an inter-point cross attention layer that uses only the invariant features of the point clouds, and an equivariant inter-point cross-attention layer described in Section 4.2.

- **Fine point matching**: As discussed in Section 4.3, we extract fine point matches between the local neighborhoods of the matched superpoints by using an optimal transport layer. We use the Sinkhorn algorithm (Sinkhorn and Knopp, 1967) for 100 steps. After extracting the soft assignment between fine points, we use solve a weighted Procrustes problem, to extract the local candidate transformations for the different matched superpoints. Finally, we follow the Local to Global Registration scheme, which selects the candidate transformation that minimizes the total alignment error.

- **Iterative Refinement:** When we perform the iterative refinement we train an initial model for the first estimation of the alignment transformation and then a second model that performs the refinement steps.

During training we supervise the output of the coarse matching module by using the overlap-aware circle loss proposed in Qin et al. (2022). Additionally, similarly to Sarlin et al. (2020) we supervise the fine point matches between the neighborhood $\mathcal{N}_{x_k}, \mathcal{N}_{y_k}$ by using a negative log-likelihood loss on the output of the soft assignment matrix Z_k produced by the optimal transport:

$$\begin{aligned}\mathcal{L}_{f,k} = -\sum_{(x,y)\in\mathcal{G}_k} \log(z_{x,y}) &- \sum_{x\in\mathcal{I}_k} \log(z_{x,m_k+1}) \\ &- \sum_{y\in\mathcal{J}_k} \log(z_{n_i+1,y})\end{aligned}$$

where $\mathcal{G}_k$ is the set of ground truth fine point matches, $\mathcal{I}_k$, $\mathcal{J}_k$ are the sets containing the rest unmatched points and $z_{.,m_k+1}$, $z_{n_i+1,.}$ corresponds to the dustbin row and column output from the learnable optimal transport module. We train our model for 40 epochs, using an initial learning rate of 10^{-4} that we reduce by a scale of 0.95 each epoch. All the parameters are optimized using the Adam optimizer (Kingma and Ba, 2015).

7.7. Local to Global Registration

The final alignment transformation is computed using a local-to-global registration scheme proposed in Qin et al. (2022). For each candidate coarse match $(x_{k(n)}, y_{k(n)})$ and their given set of inliers M_k, we compute a candidate transformation R_i, T_i by solving the optimization problem:

$$\min_{R,T} \sum_{(p,q)\in M_k} z_{p,q}\|Rp + t - q\|_2^2$$

where $z_{p,q}$ is the entry corresponding to the soft assignment of the fine point p to the point q in the optimal transport matrix Z_k. Finally we pick as the global estimated transformation, the candidate that minimizes the alignment error over the combined set of inliers $\bigcup_{k=1,\ldots,M} M_k$.

7.8. Evaluation Metrics

Registration Recall (RR): the fraction of point clouds whose estimated transformation has an error less by a set threshold. Specifically given a ground truth transformation P_{gt} and the estimated transformation P_{est} we compute the RMSE error:

$$\text{RMSE} = \sqrt{\frac{1}{|Y|}\sum_{y \in Y} \|P_{gt}^{-1} P_{est} y - y\|_2^2}$$

then the registration recall counts the fraction of registration with RMSE$<$ 0.2m.

Inlier Ratio (IR) the fraction of fine point correspondences where their residual under the ground-truth transformation is below 0.1m.

Relative Rotation and Relative Translation Error: the relative rotation error and relative translation error between the estimated and ground truth transformation

7.9. Qualitative Results

In Figure 4 we provide additional qualitative results with registrations achieved by our method. We show examples of both high and low overlap from the test set of 3DMatch and 3DLoMatch.

7.10. Limitations

One limitation of the current network is that, while in the robust setting, it achieves state-of-the-art results, in the canonical setting there is a performance gap with the current best methods. We conjecture that this can be attributed to the feature extraction backbone VNN Deng et al. (2021) and we will investigate alternatives in the future.

Another limitation of the pipeline is an additional memory overhead coming from the tensor products in the attention modules. While we did our best to create a scalable and compact architecture, the toll to satisfy the equivariance constraint exactly is that some blocks might require additional operations to their non-equivariant counterparts. While in the 3DMatch setting, this did not make a difference, the method has to be adapted properly in order to register scenes with millions of points.

A general limitation of correspondence-based methods like ours is that when the overlap is zero as in Point Cloud Assembly tasks the network cannot treat PCR properly. Moreover, as typical in PCR literature, it is implicitly assumed that there is a correct alignment for the input pairs. The network is designed to predict the best alignment possible even when no alignment is correct. Thus in order to integrate it into bigger SLAM pipelines for loop closure detection etc. additional extensions need to be done.

Lastly, the case of symmetric parts where multiple alignments are possible is not treated in this work. However, we conjecture that the advantages of our method in equivariant feature extraction from the neighboorhoods together with the local robust estimators (LGR) that propose different rotations per-neighboorhood before selecting a single one can lead to multiple consistent hypotheses in the cases of symmetric objects.

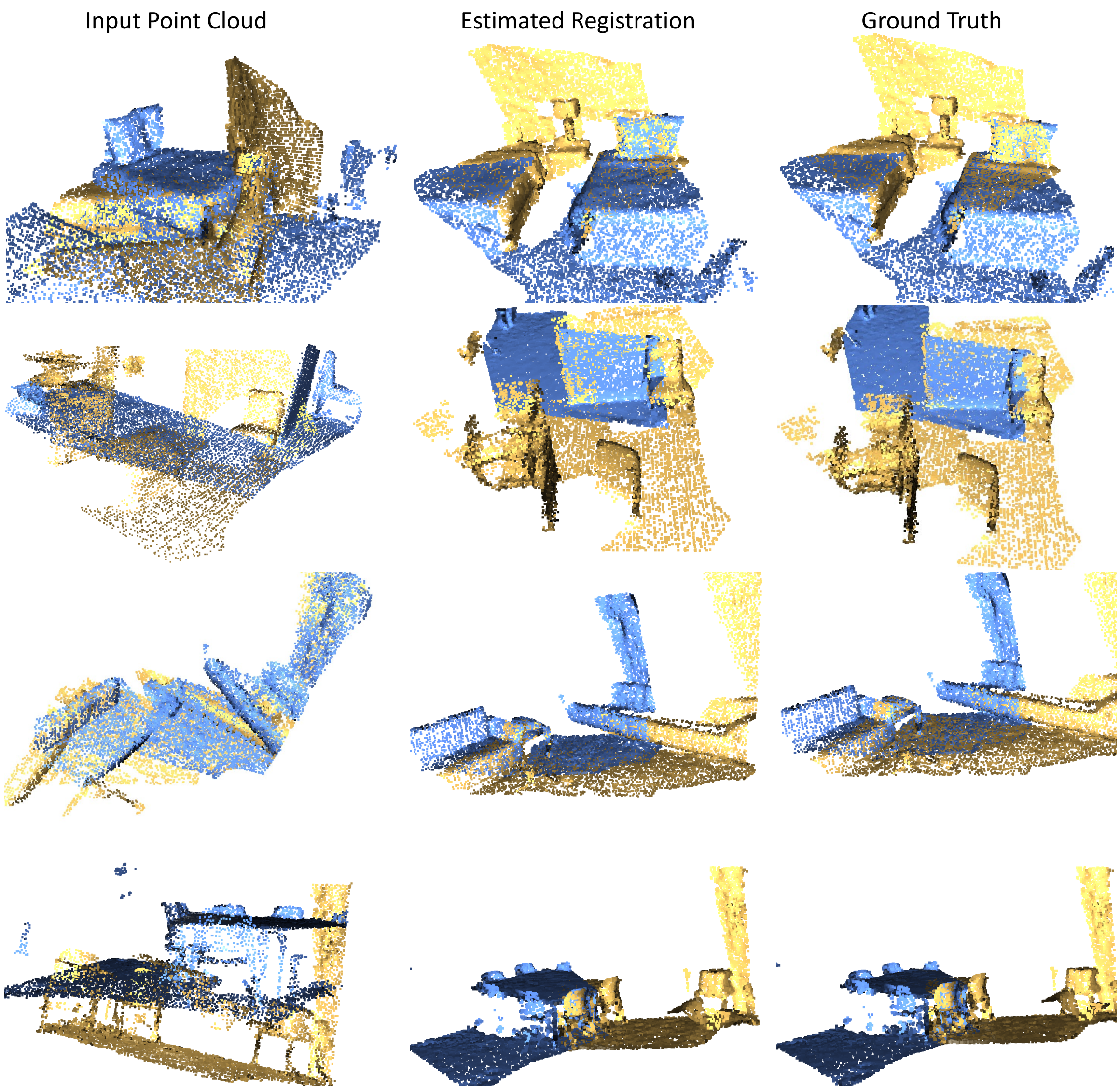

Figure 4: Registration results achieved by our method compared to the ground truth alignment.

7.11. Broader Impact

In this work, we address a major robustness limitation of current deep learning methods on point cloud registration. Our theoretical and methodological contributions, for example the novel bi-equivariant layers presented, have the potential to advance any pipeline that respects similar symmetries (for example pick-and-place in robotics manipulation).

Moreover, Point Cloud Registration can be used as the front end of larger SLAM pipelines. Our method guarantees that the registration will be consistent w.r.t. the scan poses meaning that there is no adversarial pose that would make the network behave erratically. If PCR is integrated into safety-critical applications this is a major advancement on verifiable safety.