Extreme Relative Pose Network under Hybrid Representations

Zhenpei Yang∗
UT Austin

Siming Yan∗
UT Austin

Qixing Huang†
UT Austin

Abstract

In this paper, we introduce a novel RGB-D based relative pose estimation approach that is suitable for small-overlapping or non-overlapping scans and can output multiple relative poses. Our method performs scene completion and matches the completed scans. However, instead of using a fixed representation for completion, the key idea is to utilize hybrid representations that combine 360-image, 2D image-based layout, and planar patches. This approach offers adaptively feature representations for relative pose estimation. Besides, we introduce a global-2-local matching procedure, which utilizes initial relative poses obtained during the global phase to detect and then integrate geometric relations for pose refinement. Experimental results justify the potential of this approach across a wide range of benchmark datasets. For example, on ScanNet, the rotation translation errors of the top-1/top-5 predictions of our approach are 28.6°/0.90m and 16.8°/0.76m, respectively. Our approach also considerably boosts the performance of multi-scan reconstruction in few-view reconstruction settings.

1. Introduction

Estimating the relative pose between two RGB-D scans is a crucial problem in 3D vision and robotics. In this paper, we are interested in the setting where the overlapping region between two input scans is small, or they may not even overlap. Efficient and robust solutions to this extreme relative pose problem (c.f. [49]) enjoy a wide range of applications. Examples include 3D reconstruction from a few views [12] (e.g., RGB-D scans of an indoor scene captured at a few distinctive locations), enhancing the performance of interactive scanning [42, 35] when there are interruptions during the acquisition process, early detection of loop closure [2], and solving jigsaw puzzles [8].

A standard pipeline in relative pose estimation or object matching, in general, is to first extract features from the input objects and then match these features to derive relative poses or dense correspondences [19, 1]. While early works [15, 21] use hand-crafted features or matching proce-

∗Both authors contributed equally
†huangqx@cs.utexas.edu

Figure 1: Results of multi-scan alignment under the few-view setting. Each block shows the results from 5 randomly sampled scans of a 3D scene that show minimal overlaps. (a) RobustRecons [9]. (b) Our approach. (c) Ground-truth.
non-planar scenes). Second, the matched features are adaptive to the input scan type and the relation between the input scans (e.g., dense point-wise features among overlapping regions and primitive features for non-overlapping scans).

The second aspect lies in the type of constraints between features for relative pose estimation. Instead of using a single module, our approach follows a global-2-local procedure. This procedure makes it possible to utilize different constraints at various pipeline stages. Specifically, the global matching module computes initial relative poses by performing spectral matching [28, 20] to align features extracted from the completed scans. The local refinement module, which benefits from having initial relative poses, detects and integrates additional pairwise geometric relations (e.g., parallel planes and distances between planes) for pose refinement. This sequential procedure can be easily modified to output multiple relative poses to model ambiguities in extreme relative pose estimation.

We have evaluated our approach on scan pairs from three benchmark datasets, including SUNCG [44], Matterport [6], and ScanNet [11]. For randomly sampled scan pairs, the mean top-1 rotation/translation errors of our approach are 18.1°/0.16m, 22.3°/0.39m, and 28.6°/0.90m, on SUNCG, Matterport, and ScanNet, respectively. In contrast, the state-of-the-art approach achieved 31.1°/0.27m, 34.2°/0.53m, and 34.1°/0.93m, respectively. These improvements show the advantage of using a hybrid 3D representation for relative pose estimation. Moreover, the mean top-5 errors in rotation/translation dropped to 7.8°/0.21m, 14.3°/0.41m, and 16.8°/0.76m. These performance gains show that ambiguities are abundant in extreme relative pose estimation, and the multiple solutions returned by our approach can effectively address such ambiguities.

Besides, we demonstrate the usefulness of our approach in multi-scan reconstruction. Experimental results show that combining our method and the pose optimizer of MRF-SFM [10] enables faithful 3D reconstructions from a few views and outperforms state-of-the-art multi-scan reconstruction approaches (See Figure 1).

2. Related Works

Relative pose estimation via feature matching. Global relative pose estimation approaches usually fall into two categories (c.f. [30, 48]). The first category of approaches leverages global descriptors or performs Fourier transform to align two input scans. These methods are most suitable for matching complete objects. The second category of approaches performs feature matching, e.g., RANSAC [13], robust regression [4], and spectral matching [20, 29], to estimate relative poses. However, these approaches still require that the input scans possess considerable overlapping regions. In contrast, this paper focuses on relative pose estimation between potentially non-overlapping scans.

Relative pose estimation via local optimization. Besides global matching, another category of relative pose estimation approaches focuses on local pose refinement [46]. A standard method is geometric alignment [3, 7], which minimizes point-wise distances between a pair of scans in their overlapping region. However, existing approaches typically require that the input scans are overlapping (c.f. [46]). In contrast, this paper proposes to solve the local pose refinement problem by learning and enforcing geometric relation (perpendicular, parallel, co-planar), which is suitable for both overlapping and non-overlapping scans.

Learning based relative pose estimation. Thanks to advances in deep learning that provide powerful tools to establish maps between different domains, recent works [33, 47, 51, 24, 39] formulate relative pose estimation as training pose estimation networks. Early works [33, 47, 51] usually follow the procedure of feature extraction and then relative pose estimation using a correlation module. Recent works [24, 39] employ more sophisticated matching techniques such as recurrent neural networks. However, most existing approaches still require that the two input scans possess significant overlaps, with the noticeable exception of [31], which first estimate local layout for partial scans then use hand-coded floorplane priors to assemble all pieces. Our approach differs in that we leverage the priors by explicitly learning a hybrid completion components from data.

This work is most relevant to [49], which presents a relative pose estimation approach that first performs scene completion and then computes a single relative pose prediction via geometric matching on the completed scenes. The difference in this work is that we employ a hybrid 3D scene completion in contrast to the 360-image employed in [49]. Besides, our approach introduces a novel geometric matching procedure that outputs multiple solutions and a learned local pose refinement module.

Scene completion. Our approach is also motivated by recent advances in inferring complete environments from partial observations [37, 23, 45, 22, 52]. However, our approach differs from these approaches in two ways. First, in contrast to returning the completion as the final output [45, 52] or utilizing it for learning feature representations [37, 22] or motion policies [23], our approach treats completions as an intermediate representation for feature extraction and detecting geometric relations. More importantly, instead of using a single representation for scene completion, our approach utilizes multiple representations and lets the geometric matching module to pick the most suitable representations for relative pose estimation.

3. Approach

This section describes the technical details of our approach. Section 3.1 presents an overview of our approach. Section 3.2 to Section 3.4 discuss each component in depth.
3.1. Problem Statement and Approach Overview

Our approach takes two RGB-D scans \( S_1 \) and \( S_2 \) of the same scene as input and outputs multiple rigid transformations \( T_{12} = \{ T_{12} \} \subset SE(3) \) that align the two input scans. As illustrated in Figure 2, our approach combines a global module, which generates a set of candidate relative poses, and a local module, which refines each candidate relative pose. In doing so, our approach can output multiple relative poses.

Global module. This module follows the standard pipeline of first extracting features from the input scans and then matching the extracted features. It begins with performing scan completion to generate sufficient shared features between small-overlapping or non-overlapping scans. In particular, our approach utilizes three intermediate representations, i.e., 360-image \([49]\), 2D layout \([40]\), and planar primitives \([32]\), for scan completion. Given the completed scans, this module then establishes consistent correspondences between extracted features and fit a rigid transformation to each consistent set. To this end, we apply spectral matching, which extracts consistent subsets of feature correspondences by computing leading eigenvectors of a consistency matrix among candidate correspondences. Each resulting correspondence set generates an initial relative pose by rigid pose regression.

Local module. In the same spirit as using ICP \([3]\) to optimize relative poses, our approach employs a local module to refine each output of the global module. The motivation comes from the fact initial relative poses enable additional constraints, such as iteratively computed dense correspondences and geometric relations (i.e., co-planar, parallel planes, and perpendicular planes), for relative pose estimation. Our approach also employs robust norms to remove outliers in dense correspondences and geometric relations.

Network training. A fundamental property of our approach is that the local module solves a robust optimization problem. Its optimal solution is generally insensitive to small perturbations to the outputs of the global module. Thus, we introduce two objectives for network training; namely, the optimal solution of the local module is close to the underlying ground-truth, and the global module provides effective initial solutions. The total objective function for network training consists of three terms. The first term trains the completion networks. The second term jointly trains the completion networks and the spectral matching procedure. The third term is enforced on the output of the local module.

3.2. Global Module

The global module computes a set of relative poses \( T = \{ (R_{12}, T_{12}) \} \) between the input scans \( S_1 \) with \( S_2 \). This module consists of two sub-modules. The first sub-module performs scan completion and feature extraction from each input scan \( S_i \). The result is summarized into a collection of features \( F_\theta (S_i) = \{ f \} \), where \( \theta \) denotes network parameters. Each \( f = \{ p_f, n_f, d_f \} \) consists of a 3D position \( p_f \), a 3D normal \( n_f \), and a feature descriptor \( d_f \). Note that although all extracted features share the same encoding, this sub-module combines the strengths of scan completion networks that are available under multiple representations (e.g., 360-image, 2D-layout, and 3D planar patches).

The second sub-module computes initial relative poses from correspondences between these extracted features. Since the local module will refine each relative pose, this sub-module uses spectral matching \([28, 20]\), which is efficient and can output multiple relative poses.

Scan completion sub-module. Our approach considers three data representations (See Figure 3). Since we primarily use state-of-the-art scan completion results, we only highlight the main characteristics of each completion network and defer the details to the supp. material.

The first representation is 360-image \([45, 49]\), which represents a given 3D scene as a collection of multi-channel images. The same as \([49]\), we encode for each pixel its position \( p_f \), normal \( n_f \). The completion network admits an encoder-decoder architecture and outputs predicted position, normal, and descriptor channels. This representation can generate dense completions. Besides, it is a generic representation that does not place any assumptions about
the underlying scene. However, the quality of the predicted features drops substantially among regions that are far from the visible area of each scan (c.f. [49]).

The second representation is 2D layout. We first to predict the plane of the floor of each scan. We then turn each input scan into an top-down view image by projecting points of the scan onto a grid representation of the floor. Similar to [27], the feature descriptor of each pixel collects the averaged feature among 4 vertical bins. We train the feature descriptor jointly with contrastive loss same as [49] and semantic segmentation loss. The completion network again employs an encoder-decoder architecture.

The third representation is planar patches [45, 43, 32]. To generate this representation, we perform RANSAC [14] to extract planar patches from each input scan and the corresponding complete 3D scene. We then perform PCA on the points associated with each patch to extract its patch center. The position \( p_f \) and normal \( n_f \) of this planar feature are simply the patch center and plane normal, respectively. Its feature descriptor \( d_f \) is the mean feature descriptor among the points associated with this planar patch. The completion network employs a variant of PointNet [38] by treating each plane as a generalized point. Experimentally, we found that this planar patch representation leads to better generalization performance than the first two representations, but it only involves planar patches.

**Spectral matching sub-module.** The output of the completion sub-module is a feature set \( \mathcal{F}_{\theta_g}(S_i) \) for each scan \( S_i \) (that include dense pixels and planar features). Let \( \mathcal{M}_{\theta_g,p} \subset \mathcal{F}_{\theta_g}(S_1) \times \mathcal{F}_{\theta_g}(S_2) \) collect all correspondences between the same type of features for the scan pair \( p = (S_1, S_2) \). For representations of 360-image and 2D layout, we match SIFT keypoint among the visible region of one scan to the closest point on the other scan (in terms of feature descriptors) to reduce the number of feature correspondences. As the global module only provides initial solutions for the local module, we found that increasing the number of feature matches has minor impacts on the results of the local module.

The spectral matching sub-module extracts consistent correspondence sets \( \mathcal{M}_k \subset \mathcal{M}_{\theta_g,p}, 1 \leq k \leq K \) (\( K = 5 \) in our experiments). Each correspondence set generates one initial relative pose. This step follows [28, 49], which construct a consistency matrix \( C_{\theta_g,\alpha_g,p} \in \mathbb{R}^{[\mathcal{M}_{\theta_g,p}] \times [\mathcal{M}_{\theta_g,p}]} \). The value of each element

\[
C_{\theta_g,\alpha_g,p}(c, c') = s_c s_{c'} g_{c, c'}
\]

where \( c \) and \( c' \) represent a pair of correspondences, \( s_c \) models the descriptor similarity of \( c \), \( g_{c, c'} \) measures the preservation of distances and angles between \( c \) and \( c' \), and \( \alpha_g \) represent hyper-parameters of \( s_c \) and \( g_{c, c'} \). Due to the space constraint, please refer to the supp. material for details.

Given \( C_{\theta_g,\alpha_g,p} \), we follow [28, 49] to extract top K matches, each of which is given by an indicator vector \( u_k \in \{0, 1\}^{[\mathcal{M}_{\theta_g,p}]} \). We employ the rigid regression formulation in [18] to obtain the initial relative pose \( R_k, t_k \):

\[
\min_{R, t} \sum_{c=(f_1, f_2)} u_k(c)||R f_1 + t - p_{f_2}||^2
\]

### 3.3. Local Module

The local module refines each initial relative pose obtained from the global module. Benefited from having initial relative poses, this module employs different sets of features for refinement. Specifically, it performs robust regression on iteratively updated closest point pairs and predicted geometric relations (e.g., point pairs that lie on the same plane) between the completed scans. In the following, we first introduce how to predict geometric relations. We then describe how to perform robust regression.

**Geometric relations.** We consider three types of geometric relations that frequently occur in man-made scenes: point pairs that share the same plane, point pairs that lie on perpendicular planes, and point pairs that lie on parallel planes. Consider a down-sampled set of features \( \mathcal{F}_{\theta_g}(S_i) \subset \mathcal{F}_{\theta_g}(S_i) \) for each scan \( S_i \). Our implementation uses 6000 random samples for each scan pair. Given all feature pairs \( \mathcal{M}_{\theta_g,p} \) between these features the local module trains a variant of PointNet++ [38] to classify each pair \( c \in \mathcal{M}_{\theta_g,p} \) into either no relation or one of the relations described above, i.e., four categories in total. In the following, we denote the probability of \( c \) belonging to the categories of co-planar, parallel, and perpendicular as \( w_{\theta_i}(c) \in [0, 1] \), respectively, where \( \theta_i \) denotes the network parameters of the relation prediction network.

**Robust regression.** We begin with defining consistency scores between a predicted geometric relation \( c = (f_1, f_2) \) and the current rigid transformation \( (R, t) \):

\[
f^1_R(c) := \frac{||p_{f_1} + t - p_{f_2}||^2}{||p_{f_1}||^2 + ||p_{f_2}||^2}
\]

\[
f^2_R(c) := 1 - \frac{||R n_{f_1}||^2}{||n_{f_1}||^2}
\]

\[
f^3_R(c) := \frac{||R n_{f_1}||^2}{||n_{f_1}||^2}
\]
In other words, \( f_1, f_2, \) and \( f_3 \) force a point pair to satisfy the co-planar, parallel, and perpendicular relations, respectively. In addition, \( f_1 \) also serves as a symmetric point-to-plane measure for matching closest point pairs (c.f. [41]).

The robust regression formulation generalizes the standard approach of using iteratively reweighted non-linear least squares for rigid alignment [46]. Specifically, at each iteration, it solves the following non-linear least squares:

\[
\min_{\hat{R}, \hat{t}} \sum_{c \in M_{\theta_{\hat{R}, \hat{t}}}} w_0^c f_{\theta_{\hat{R}, \hat{t}}}(c) + \alpha_c \left( \sum_{c \in M_{\theta_{\hat{R}, \hat{t}}}} w_1^c \hat{w}_\theta(c) f_{\hat{R}, \hat{t}}(c) \right.
+ \sum_{c \in M_{\theta_{\hat{R}, \hat{t}}}} w_2^c \hat{w}^2_\theta(c) f_\theta(c) + \sum_{c \in M_{\theta_{\hat{R}, \hat{t}}}} w_3^c \hat{w}^3_\theta(c) f_\theta^3(c) \right) \tag{5}
\]

where \( M_{\theta_{\hat{R}, \hat{t}}} \) collects nearest neighbor pairs from features of \( \mathcal{F}_{\theta_{\hat{R}}}(S_1) \) to \( \mathcal{F}_{\theta_{\hat{t}}}(S_2) \), and vice-versa, based on the relative pose at the previous iteration. \( w_0^c, w_1^c, w_2^c, \) and \( w_3^c \) are updated adaptively based on the current relative pose \( \hat{R}, \hat{t} \):

\[
w_0^c := \alpha_0^2/(\alpha_0^2 + f_{\theta_{\hat{R}, \hat{t}}}(c)) \quad \quad w_1^c := \alpha_1^2/(\alpha_1^2 + f_{\hat{R}, \hat{t}}(c)) \quad \quad w_2^c := \alpha_2^2/(\alpha_2^2 + f_\theta^2(c)) \quad \quad w_3^c := \alpha_3^2/(\alpha_3^2 + f_\theta^3(c)) \tag{6}
\]

where \( \alpha_i = \{\alpha_r, \alpha_0, \alpha_1, \alpha_2, \alpha_3\} \) are hyper-parameters, which will be trained together with \( \theta_\phi \) and \( \theta_1 \).

With this setup, the robust regression procedure first applies (6) to set the term weights using the initial relative pose. Then it alternates among computing the nearest neighbors, applying the Gauss-Newton method to solve (5), and updating the term weights. In our implementation, we run 3 alternating iterations.

### 3.4. Network Training

Since the local module solves a robust optimization problem, its optimal solution is insensitive to small perturbations to the output of the global module. Thus, instead of training the combination of the local module and the global module end-to-end, we utilize a decoupled approach in this paper. Specifically, our approach optimizes all variables \( \Theta = \{\theta_\phi, \alpha_\phi, \theta_1, \alpha_1\} \) by minimizing the following loss term:

\[
\min_{\Theta} l_4(\theta_{\phi}) + \lambda_2 l_2(\theta_1) + \lambda_3 l_3(\alpha_{\phi}) + \lambda_4 l_4(\alpha_{\phi}) \tag{7}
\]

where \( l_1, l_2, l_3, \) and \( l_4 \) train scan completion networks, the geometric relation network, the spectral matching submodule, and the local module, respectively. \( \lambda_i \) are hyper-parameters determined via 10-fold cross validation.

The scan completion loss \( l_1(\theta_{\phi}) \) combines the losses for training the scan completion network under each data representation. The geometric relation loss \( l_2(\theta_1) \) utilizes initial relative poses that randomly perturbed from the underlying ground-truth. Since the formulations of these two losses are standard for network training, we leave the details to the supp. material. In the following, we focus on the loss terms for spectral matching sub-module and the local module.

#### Loss term for spectral matching

Consider the consistency matrix \( C_p \) for a scan pair \( p = (S_1, S_2) \). Our goal for training the spectral matching sub-module is to enforce that the normalized indicator vector \( u_p^{gt} \) of ground-truth correspondences is a top eigenvector of \( C_p \). Note that \( C_p \) is dependent on \( \alpha_\phi \). We omit them to make the notations uncluttered.

To avoid training a recurrent neural network converted from power iteration for eigen-vector computation (c.f. [36]), our approach utilizes three properties of a top eigen-vector, i.e., a normalized vector \( u \) is a top eigenvector of a matrix \( C \) if (1) the corresponding eigenvalue \( \lambda = u^T C u \), (2) \( C u = \lambda u \), and (3) \( \lambda \) is maximized. Given a collection of scan pairs \( P = \{p\} \) as, we define \( l_4(\alpha_{\phi}) := \)

\[
\sum_{p \in P} (\| C_p u_p^{gt} - (u_p^{gt}^T C_p u_p^{gt}) u_p^{gt} \|^2 - u_p^{gtT} C_p u_p^{gtT})
\]

#### Loss term for the local module

Define \( I_p(R, t) \in \mathbb{R}^{3 \times 4} \) as one iteration of the IRNLS procedure from the current pose \( (R, t) \), i.e., optimization and reweighting. Similar to \( l_3 \), we avoid training recurrent networks of the robust regression procedure by directly penalizing the difference between the ground-truth relative pose \( \hat{R}_p, \hat{t}_p \) and \( I_p(R_p, t_p) \):

\[
l_4(\alpha_{\phi}) = \sum_{p \in P} \| (R_p, t_p) - I_p(R_p^{gt}, t_p^{gt}) \|_F^2 \tag{8}
\]

where \( \| \cdot \|_F \) is the matrix Frobenius-norm. One intuition of (8) is that when \( (R_p^{gt}, t_p^{gt}) = I_p(R_p^{gt}, t_p^{gt}) \), then \( (R_p, t_p) \) becomes the stationary point of the IRNLS procedure.

Minimizing \( l_4(\alpha_{\phi}) \) requires computing the derivatives of \( I_p \). Since \( I_p \) is based on a local optimum of (5), we use the implicit function theorem [26] to back-propagate the gradients from the optimized poses to the predicted features and geometric relations. Again, we defer the details to the supp. material for the interest of space.

#### Training procedure

We train (7) through three phases. The first phase trains each prediction network in isolation. The second phase fixes the prediction networks and trains the hyper-parameters of each module in isolation. The final phase fine-tunes all the variables together by solving (7). Network training involves the ADAM optimizer [25], and please refer to the supp. material for details.

### 4. Experimental Evaluation

This section presents an experimental evaluation of the proposed approach. Section 4.1 describes the experimental setup. Section 4.2 analyzes the experimental results and compares the proposed approach with baseline approaches. Section 4.3 and Section 4.4 analyze the global module and the local module, respectively.
Figure 4: Qualitative comparison between our approach and baseline approaches. From left to right, we show the results of Super4PCS [34], RobustGR [50], ScanComplete [49], Ours-Top1, Ours-Top3, and Ground Truth figures.

4.1. Experimental Setup

Datasets. We perform experimental evaluation using three datasets: SUNCG [44], Matterport [6], and ScanNet [11], where SUNCG provides synthetic data, Matterport and ScanNet provide real data. For each room in each dataset, we randomly sample 25 camera locations. For real data sets such as ScanNet, the camera locations are sampled from the recording sequences. For each dataset, we collect around 50k scan pairs for training, and 1k scan pairs for testing. Note that the scenes of the training and testing scans do not overlap.

Baseline approaches. We consider the five baseline approaches: Super4PCS [34], RobustGR [50], ScanComplete [49], Ours-Global, and Ours-Local. Here, Super4PCS is a widely used global scan matching method; RobustGR is a state-of-the-art global scan matching method that is based on robust regression; ScanComplete is a state-of-the-art approach for indoor RGB-D registration that performs scan completion using the 360-image representation. Moreover, Ours-Global is our approach without the local module. Ours-Local is our approach without the global module, i.e., applying the local module directly on the completed scans.

Evaluation protocol. We follow the standard protocol of reporting the relative angular error $\frac{||\log(R^* R_{gtT})||}{\sqrt{2}}$ and the relative translation error $||t^* - t_{gt}||$. Here $(R_{gt}, t_{gt})$ and $(R^*, t^*)$ denote the ground-truth and the output of a method, respectively. In the presence of multiple outputs, we define the best match as the one that minimizes the rotation error. In this section, we report the rotation and the translation errors among the top 1, 3, and 5 matches. The order of the matches are given by the eigenvalues of the consistency matrix $C_{\theta, \alpha, p}$ and the final scores of (5), respectively. In addition, we divide scan pairs into two categories: overlapping and non-overlapping. The category of overlapping scans consists of scans pairs that overlap by at least 10% points (with respect to the scan with fewer num-
<table>
<thead>
<tr>
<th></th>
<th>SUNCG</th>
<th>Matterport</th>
<th>ScanNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rotation Trans</td>
<td>Rotation Trans</td>
<td>Rotation Trans</td>
</tr>
<tr>
<td>4PCS-Overlap (≥10%)</td>
<td>45.2 0.22</td>
<td>49.7 0.55</td>
<td>31.3 0.94</td>
</tr>
<tr>
<td>RobustGR (≥10%)</td>
<td>64.9 0.57</td>
<td>38.1 0.69</td>
<td>45.4 1.20</td>
</tr>
<tr>
<td>ScanComp. (≥10%)</td>
<td>9.8 0.17</td>
<td>19.9 0.39</td>
<td>18.4 0.71</td>
</tr>
<tr>
<td>Ours-Local (≥10%)</td>
<td>56.0 0.22</td>
<td>40.8 0.49</td>
<td>25.4 0.93</td>
</tr>
<tr>
<td>Ours-GLOBAL (≥10%)</td>
<td>8.0 0.16</td>
<td>13.9 0.31</td>
<td>16.7 0.69</td>
</tr>
<tr>
<td>Ours (≥10%)</td>
<td>5.0 0.11</td>
<td>10.7 0.29</td>
<td>14.1 0.67</td>
</tr>
<tr>
<td>ScanComp. (≤10%)</td>
<td>79.7 0.49</td>
<td>70.0 0.88</td>
<td>66.4 1.37</td>
</tr>
<tr>
<td>Ours-Local (≤10%)</td>
<td>105.0 0.31</td>
<td>115.6 0.49</td>
<td>67.2 1.34</td>
</tr>
<tr>
<td>Ours-GLOBAL (≤10%)</td>
<td>62.8 0.40</td>
<td>59.3 0.69</td>
<td>59.3 1.36</td>
</tr>
<tr>
<td>Ours (≤10%)</td>
<td>60.9 0.41</td>
<td>54.1 0.68</td>
<td>58.4 1.36</td>
</tr>
<tr>
<td>ScanComp. (all)</td>
<td>31.1 0.27</td>
<td>34.2 0.53</td>
<td>34.1 0.93</td>
</tr>
<tr>
<td>Ours-Local (all)</td>
<td>56.9 0.24</td>
<td>60.7 0.49</td>
<td>36.6 0.97</td>
</tr>
<tr>
<td>Ours-GLOBAL (all)</td>
<td>2.4 0.23</td>
<td>26.0 0.41</td>
<td>30.7 0.91</td>
</tr>
<tr>
<td>Ours (all)</td>
<td>18.1 0.16</td>
<td>22.3 0.39</td>
<td>28.6 0.90</td>
</tr>
</tbody>
</table>

Table 2: Benchmark evaluation on our approach and baseline approaches. Ours-Local and Ours-GLOBAL stand for our method with global module and local module removed, respectively. We show the mean error for rotation and translation components for overlapping scan pairs (≥10%) and non-overlapping scan pairs (≤10%), respectively.

4.2. Analysis of Results

Table 2 and Figure 4 provide quantitative and qualitative results of our approach and baseline methods. Overall, our approach outperforms baseline approaches considerably. The mean rotation/translation errors of the best-match of our approach are 18.1°/0.16m, 22.3°/0.39m, and 28.6°/0.90m on SUNCG, Matterport, and ScanNet, respectively. In contrast, the state of the art method only achieved 31.1°/0.27m, 34.2°/0.53m, and 34.1°/0.93m, respectively.

Overlapping vs. non-overlapping. In the overlapping regime, the relative improvements of mean rotation/translation errors are 40.8%/35.3%, 46.2%/25.6%, and 55.0%/5.6% on the three datasets described above, respectively. Meanwhile, the relative improvements in the non-overlapping regime are 23.6%/16.3%, 22.9%/22.7%, and 10.5%/0.7%, respectively. These statistics show the consistency of our approach across different overlapping rates.

4.3. Analysis of the Global Module

We focus on analyzing two important aspects of the global module, namely the effects of using hybrid representation and multiple outputs.

Single representation versus hybrid representation. As shown in Table 1, using hybrid representations leads to noticeable performance gains when compared to using a single representation. For mean rotation/translation errors, the relative improvements from the top performing single representation (ScanComplete in this case) are 20.4%/14.8%, 24.0%/22.6%, and 11.7%/12% on SUNCG, Matterport, and ScanNet, respectively. It should be noted that the 2D layout and Plane representations individually do not provide sufficient features for relative pose estimation. However, they provide complementary features to the 360-image representation that boost the overall performance of our approach. This is also consistent with Figure 3, which shows that the matched features consist of a mixture of planar features and point features from 360-image and 2D layout.

Best match versus K-best. With multiple outputs, the top-5 rotation/translation errors of the global module drop significantly. The relative improvements on the three datasets are 68.4%/8.7%, 45.0%/0.1%, and 45.2%/16.5%, respectively. Besides, the relative improvements on non-overlapping scans are significantly bigger than those on overlapping scans. An explanation is that indoor scenes exhibit approximate discrete symmetries, e.g., rotational symmetries of a box-shape, and such symmetries incur ambiguities for non-overlapping scans (i.e., they lack detailed geometric features to lock their relative poses). On the other hand, the orders of these symmetries are relatively small, and top matches tend to capture them.
Table 3: Benchmark evaluation on the different combination of global initialization and local refinement. We use the strongest baseline [49] to serve as global module, and SparseICP [5] as baseline local module. We show the mean rotation and translation errors on each dataset.

<table>
<thead>
<tr>
<th></th>
<th>SUNCG</th>
<th>Matterport</th>
<th>ScanNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rot</td>
<td>Trans</td>
<td>Rot</td>
</tr>
<tr>
<td>Base-Global+Base-Local</td>
<td>31.9</td>
<td>0.45</td>
<td>33.6</td>
</tr>
<tr>
<td>Base-Global+Ours-Local</td>
<td>27.8</td>
<td>0.24</td>
<td>30.1</td>
</tr>
<tr>
<td>Ours-Global+Base-Local</td>
<td>25.9</td>
<td>0.44</td>
<td>25.4</td>
</tr>
<tr>
<td>Ours-Global+Ours-Local</td>
<td>18.1</td>
<td>0.16</td>
<td>22.3</td>
</tr>
</tbody>
</table>

In addition, the probability distribution of the rank of the matches with the best rotational error are 42%, 17%, 10%, and 14%, respectively. In other words, 58% of the best matches are not derived from the first eigen-vector. This indicates the importance of leveraging multiple outputs. In Section 4.5, we show that using multiple outputs for multi-scan reconstruction is superior to using a single output.

### 4.4. Analysis of the Local Module

Overall, the local module leads to noticeable performance gains from the outputs of the global module. As shown in Table 2, the mean rotation/translation errors of the top 1 match reduced from 24.7°/0.23m to 18.1°/0.16m on SUNCG, from 26.0°/0.41m to 22.3°/0.39m on Matterport, and from 30.7°/0.91m to 28.6°/0.90m on ScanNet, respectively. These statistics indicate the robustness of our local module.

**Baseline comparison.** Our approach outperforms the robust version of ICP [5] when applied on the output of the global module (See Table 3). The mean rotation/translation errors reduce from 25.9°/0.44m, 25.4°/0.40m, and 30.4°/0.91m to 18.1°/0.16m, 22.3°/0.34m, and 28.6°/0.90m. Such improvements mainly come from the predicted geometric relations. Our approach also outperforms the baseline that combines the output of ScanComplete and our local module. The relative improvements are salient across all the datasets. These statistics indicate that having good initial poses for local refinement is key to the success of relative pose estimation.

**The quality of geometric relations prediction.** The average accuracy of geometric relations across all three datasets are 46%, 69%, 75%, and 93% for no-relation, perpendicular, parallel, co-planar respectively. The statistics are collected from scan pairs perturbed from the underlying ground-truth by random rotation with degree in $[0, 30^\circ]$ and a random translation in $[-0.5m, 0.5m]^3$. Although the predictions are imperfect, the success of our approach lies in using robust norms to filter out wrong predictions automatically. Due to space constraints, we leave a detailed analysis of the predicted geometric relations to the supp. material.

Table 4: Comparisons on multi-scan alignment using our approach + MRF-SFM[10] and two baseline approaches. Our approach shows a clear advantage in the sparse setting, e.g., using 5 scans.

<table>
<thead>
<tr>
<th></th>
<th>5 Scans</th>
<th>10 Scans</th>
<th>15 Scans</th>
<th>20 Scans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rot</td>
<td>Trans</td>
<td>Rot</td>
<td>Trans</td>
</tr>
<tr>
<td>Fine2C[17]</td>
<td>96.5</td>
<td>1.41</td>
<td>62.4</td>
<td>0.93</td>
</tr>
<tr>
<td>RobustRec[9]</td>
<td>33.1</td>
<td>0.78</td>
<td>25.1</td>
<td>0.52</td>
</tr>
<tr>
<td>Our-top-1</td>
<td>8.39</td>
<td>0.20</td>
<td>3.55</td>
<td>0.15</td>
</tr>
<tr>
<td>Our-top-3</td>
<td>5.77</td>
<td>0.18</td>
<td>3.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Our-top-5</td>
<td>5.13</td>
<td>0.18</td>
<td>2.73</td>
<td>0.12</td>
</tr>
</tbody>
</table>

### 4.5. Application in Few-View Reconstruction

We have studied the impacts of our relative pose estimation approach in multi-scan alignment. To this end, we pair our approach with the multi-view pose optimizer of MRF-SFM [10]. The goal is to find the most consistent subset of relative poses among the outputs of our approach. To analyze the benefits of having multiple relative pose estimations between each scan pair, we report the performance of using top-1, top-3, top-5 outputs of our approach, respectively. We performed an experimental study on ScanNet, using 5/10/15/20 randomly sampled scans. Regarding the evaluation protocol, we report absolute errors in rotations and translations of aligned scans using the same metric for evaluating relative poses. Since MRF-SFM extracts consistent relative transformations, the mean rotation/translation errors of the resulting absolute transformations are significantly reduced. The baseline approaches include RobustRecons [9] and Fine-to-Coarse [17], which are two state-of-the-art multi-scan alignment approaches. Fine-to-Coarse additionally utilizes the scanning order among the input scans.

As shown in Figure 1 and Table 4, our approach significantly outperforms baseline approaches. The improvements are salient in the sparse setting, where our approach shows high-quality outputs. In such a sparse setting, RobustRecons also outperforms Fine-to-Coarse, as the latter has to utilize dense inputs to detect initial planes. Utilizing top-3 and top-5 matches also improves the performance. One explanation is that there are more correct inputs for multi-scan alignment when utilizing top-3 and top-5 matches.

### 5. Conclusions and Future Work

In this paper, we have introduced an approach for estimating the relative pose between two potentially non-overlapping RGB-D scans. Our approach combines a global module for computing candidate relative poses and a local module for refining each candidate pose. Our method can output multiple relative poses, and the resulting pose quality is considerably better than state-of-the-art approaches for both overlapping and non-overlapping scans.
References


