# Extreme Relative Pose Estimation for RGB-D Scans via Scene Completion

Zhenpei Yang
UT Austin

Jeffrey Z.Pan
Phillips Academy Andover

Linjie Luo
ByteDance AI Lab

Xiaowei Zhou
Zhejiang University*

Kristen Grauman
UT Austin

Qixing Huang
UT Austin†

## Abstract

*Estimating the relative rigid pose between two RGB-D scans of the same underlying environment is a fundamental problem in computer vision, robotics, and computer graphics. Most existing approaches allow only limited relative pose changes since they require considerable overlap between the input scans. We introduce a novel approach that extends the scope to extreme relative poses, with little or even no overlap between the input scans. The key idea is to infer more complete scene information about the underlying environment and match on the completed scans. In particular, instead of only performing scene completion from each individual scan, our approach alternates between relative pose estimation and scene completion. This allows us to perform scene completion by utilizing information from both input scans at late iterations, resulting in better results for both scene completion and relative pose estimation. Experimental results on benchmark datasets show that our approach leads to considerable improvements over state-of-the-art approaches for relative pose estimation. In particular, our approach provides encouraging relative pose estimates even between non-overlapping scans.*

## 1. Introduction

Estimating the relative rigid pose between a pair of RGB-D scans is a fundamental problem in computer vision, robotics, and computer graphics with applications to systems such as 3D reconstruction [49], structure-from-motion [38], and simultaneous localization and mapping (SLAM) [41]. Most existing approaches [12, 17, 1, 30, 46] follow a three-step paradigm (c.f. [49]): feature extraction, feature matching, and rigid transform fitting with the most consistent feature correspondences. However, this paradigm requires the input RGB-D scans to have considerable overlap, in order to establish sufficient feature correspondences for matching. For input scans of extreme rela-
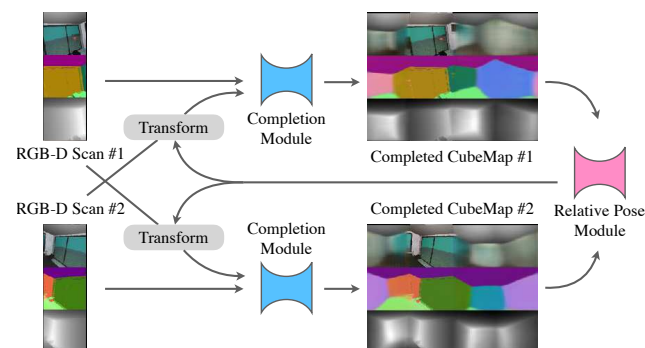


Figure 1: Illustration of the work-flow of our approach. We align two RGB-D scans by alternating between scene completion (completion module) and pose estimation (relative pose module).

tive poses with little or even *no* overlap, this paradigm falls short since there are very few or no features to be found in the overlapping regions. Nevertheless, such problem settings with minimal overlap are common in many applications such as solving jigsaw puzzles [5], early detection of loop closure for SLAM [13], and reconstruction from minimal observations, e.g., a few snapshots of an indoor environment [26].

While the conventional paradigm breaks down in this setting, we hypothesize that solutions are possible using prior knowledge for typical scene structure and object shapes. Intuitively, when humans are asked to perform pose estimation for non-overlapping inputs, they utilize the prior knowledge of the underlying geometry. For example, we can complete a human model from two non-overlapping scans of both the front and the back of a person; we can also tell the relative pose of two non-overlapping indoor scans by knowing that the layout of the room satisfies the Manhattan world assumption [7]. This suggests that when direct matching of non-overlapping scans is impossible, we seek to match them by first performing scene completions and then matching completed scans for their relative pose.

Inspired by this intuition, we introduce an approach that takes a pair of RGB-D scans with little overlap as input and outputs the relative pose between them. Key to our approach are internal modules that infer the *completion* of

---
*Xiaowei Zhou is affiliated with the StateKey Lab of CAD&CG and the ZJU-SenseTime Joint Lab of 3D Vision.
†huangqx@cs.utexas.edu

each input scan, allowing even widely separated scans to be iteratively registered with the proper relative pose via a recurrent module. As highlighted in Figure 1, our network first performs single-scan completion under a rich feature representation that combines depth, normal, and semantic descriptors. This is followed by a pair-wise matching module, which takes the current completions as input and outputs the current relative pose. To address the issue of imperfect predictions, we introduce a novel pairwise matching approach that seamlessly integrates two popular pairwise matching methodologies: spectral matching [24, 16] and robust fitting [2]. Moreover, rather than merely perform relative pose estimation from the completed input scans, we propose to alternate between scene completion and relative pose estimation. This allows us to leverage signals from both input scans to achieve better completion results in later iterations. Given progressively improved relative pose estimations, the recurrent module updates each completion accordingly by fusing information from both input scans.

Note that compared to existing deep learning methods [27, 10], the novelty of our approach is three-fold:

1. Explicitly supervising the relative pose network via completions of the underlying scene under a novel representation that combines geometry and semantics.
2. A novel pairwise matching method that combines spectral matching and iteratively reweighted least squares.
3. An iterative procedure that alternates between scene completion and pairwise matching.

We evaluate our approach on three benchmark datasets, namely, SUNCG [39], Matterport [3], and ScanNet [8]. Experimental results show that our approach is significantly better than state-of-the-art relative pose estimation techniques. For example, our approach reduces the mean rotation errors of state-of-the-art approaches from $36.6°$, $42.0°$, and $51.4°$ on SUNCG, Matterport, and ScanNet, respectively, to $12.0°$, $9.0°$, and $30.2°$, respectively, on scans with overlap ratios greater than 10%. Moreover, our approach generates encouraging results for non-overlapping scans. The mean rotation errors of our approach for these scans are $79.9°$, $87.9°$, and $81.8°$, respectively. In contrast, the expected error of a random rotation is around $126.3°$.

Code is publicly available at https://github.com/zhenpeiyang/RelativePose.

## 2. Related Work

**Non-deep learning techniques.** Pairwise object matching has been studied extensively in the literature, and it is beyond the scope of this paper to present a comprehensive overview. We refer to [21, 42, 25] for surveys on this topic and to [30, 47, 32] for recent advances. Regarding the specific task of relative pose estimation from RGB-D scans, popular methods [12, 17, 1, 30] follow a three-step procedure. The first step extracts features from each scan. The

second step establishes correspondences for the extracted features, and the third step fits a rigid transform to a subset of consistent feature correspondences. Besides the fact that the performance of these techniques heavily relies on parameter tuning for each component, they also require that the two input scans possess sufficient overlapping features to match.

**Deep learning techniques.** Recent works explore deep neural networks for the task of relative pose estimation (or pairwise matching in general) [10, 18, 43, 48, 29]. These approaches follow the standard pipeline of object matching, but they utilize a neural network module for each component. Specifically, feature extraction is generally done using a feed-forward module, while estimating correspondences and computing rigid transforms are achieved using a correlation module (c.f. [10]). With proper pre-training, these methods exhibit better performance than their non-deep learning counterparts. However, they still require that the input scans possess a sufficient overlap so that the correlation module can identify common features for relative pose estimation.

A couple of recent works propose recurrent procedures for object matching. In [36], the authors present a recurrent procedure to compute weighted correspondences for estimating the fundamental matrix between two images. In [22], the authors use recurrent networks to progressively compute dense correspondences between two images. The network design is motivated from the procedure of nonrigid image registration between a pair of images. Our approach is conceptually relevant to these approaches. However, the underlying principle for the recurrent approach is different. In particular, our approach performs scan completions, from which we compute the relative pose.

**Optimization techniques for pairwise matching.** Existing feature-based pairwise matching techniques fall into two categories. The first category of methods is based on MAP inference [24, 16, 4], where feature correspondence scores and pairwise consistency scores are integrated as unary and pairwise potentials. A popular relaxation of MAP inference is spectral relaxation [24, 16]. The second category of methods is based on fitting a rigid transformation to a set of feature correspondences [14]. In particular, state-of-the-art approaches [11, 21, 46] usually utilize robust norms to handle outlier feature correspondences. In this paper, we introduce the first approach that optimizes a single objective function to simultaneously perform spectral matching and robust regression for relative pose estimation.

**Scene completion.** Our approach is also motivated from recent advances on inferring complete environments from partial observations [34, 20, 40, 19, 50]. However, our approach differs from these approaches in two ways. First, in contrast to returning the completion as the final output [40, 50] or utilizing it for learning feature representations [34, 19] or motion policies [20], our approach treats
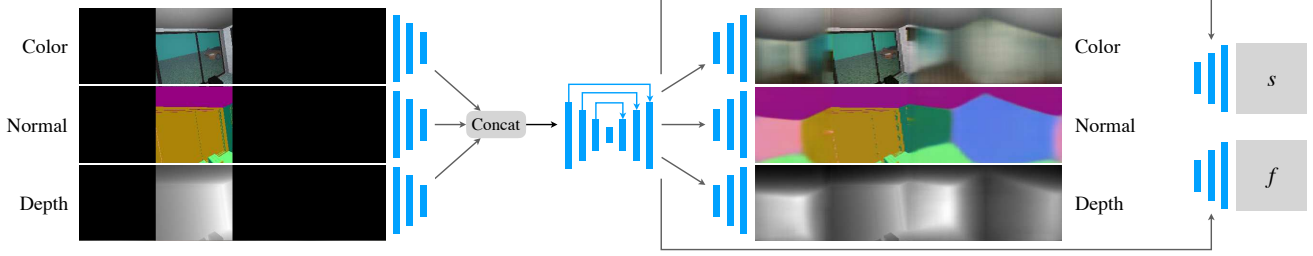
Figure 2: Network design of the completion module. Given the partially observed color, depth, normal, our network completes a cube-map representation of color, depth, normal, and semantics, as well as a feature map. Please refer to Sec. 3.3 for details.

completions as an intermediate representation for relative pose estimation. From the representation perspective, our approach predicts color, depth, normal, semantic, and feature vectors using a single network.

# 3. Approach

We begin with presenting an approach overview in Section 3.1. Section 3.2 to Section 3.4 elaborate the network design. Section 3.5 discusses the training procedure.

## 3.1. Approach Overview

The relative pose estimation problem studied in this paper considers two RGB-D scans $I_i \in \mathbb{R}^{h \times w \times 4}, 1 \leq i \leq 2$ of the same environment as input ($h,w$=160 in this paper). We assume that the intrinsic camera parameters are given so that we can extract the 3D position of each pixel in the local coordinate system of each $I_i$. The output is a rigid transformation $T = (R \in \mathbb{R}^{3 \times 3}, t \in \mathbb{R}^3) \in \mathbb{R}^{3 \times 4}$ that characterizes the relative pose between $I_1$ and $I_2$. Note that we do not assume $I_1$ and $I_2$ overlap.

Our approach is inspired from simultaneous registration and reconstruction (or SRAR) [15], which aligns input scans to a deforming surface (expressed in a world coordinate system). The key advantage of SRAR is that the deforming surface provides an complete intermediate representation for aligning scans that do not necessarily overlap. However, directly applying SRAR to relative pose estimation for 3D scenes is challenging, as unlike 3D objects [44, 35, 6, 45], it is difficult to specify a world coordinate system for 3D scenes. To address this issue, we modify SRAR by maintaining two copies $S_1$ and $S_2$ of the complete underlying environment, where $S_i$ is expressed in the local coordinate system of $I_i$ (We will discuss the precise representation of $S_i$ later.) Conceptually, our approach reconstructs each $S_i$ by combining the signals in both $I_1$ and $I_2$. When performing relative pose estimation, our approach employs $S_1$ and $S_2$, which addresses the issue of non-overlap.

As illustrated in Figure 1, the proposed network for our approach combines a scan completion module and a pairwise matching module. To provide sufficient signals for pairwise matching, we define the feature representation $\overline{X}, X \in \{I_1, I_2, S_1, S_2\}$ by concatenating color, depth, nor-

mal, semantic labels, and descriptors. Here $\overline{S}_i$ utilizes a reduced cube-map representation [40], where each face of $\overline{S}_i$ shares the same representation as $\overline{I}_i$. Experimentally, we found this approach gives far better results than performing scan completion under the RGB-D representation first and then computing the feature representation.

Under this feature representation, our approach first estimates $\overline{S}_i$ from the corresponding $\overline{I}_i$. The pairwise matching module takes current $\overline{S}_1$ and $\overline{S}_2$ as input and outputs the current relative pose $T$. The completion module updates each scan completion using the transformed scans, e.g., $\overline{S}_1$ is estimated from $\overline{I}_1$ and the transformed $\overline{I}_2$ in the local coordinate system of $I_1$. We alternate between applying the pairwise matching module and the scan completion module. In our implementation, we use three recurrent steps. In terms of network and parameter learning, we first train the scan completion networks using ground-truth completions as the supervision. We then optimize the hyper-parameters of the relative pose module to maximize the end-to-end performance of our approach. Next we elaborate on the details.

## 3.2. Feature Representation

Motivated by the particular design of our pairwise matching module, we define the feature representation of an RGB-D scan $I$ as $\overline{I} = (c, d, n, s, f)$. Here $c \in \mathbb{R}^{h \times w \times 3}$, $d \in \mathbb{R}^{h \times w \times 1}$, $n \in \mathbb{R}^{h \times w \times 3}$, $s \in \mathbb{R}^{h \times w \times n_c}$, $f \in \mathbb{R}^{h \times w \times k}$($k$=32 in this paper), specifying color, depth, normal, semantic class, and a learned descriptor, respectively. The color, depth, normal, and semantic class are obtained using the densely labeled reconstructed model for all datasets.

## 3.3. Scan Completion Modules

The scan completion module takes in a source scan, a target scan transformed by current estimate $T$ (not used for the first iteration), and outputs the complete feature representation $\overline{S}_i$. We encode $\overline{S}_i$ using a reduced cube-map representation [40], which consists of four faces (excluding the floor and the ceiling). Each face of $\overline{S}_i$ shares the same feature representation as $\overline{I}_i$. For convenience, we always write $\overline{S}_i$ in the tensor form as $\overline{S}_i = (\overline{S}_{i,1}, \overline{S}_{i,2}, \overline{S}_{i,3}, \overline{S}_{i,4}) \in \mathbb{R}^{h \times w \times 4(k+n_c+7)}$. Following the convention [40, 34], we formulate the input to both scan completion modules us-

ing a similar tensor form $\hat{I}_i = (\hat{I}_{i,1}, \hat{I}_{i,2}, \hat{I}_{i,3}, \hat{I}_{i,4}) \in \mathbb{R}^{h \times w \times 4(k+n_c+8)}$, where the last channel is a mask that indicates the presence of data. As illustrated in Figure 2 (Left), we always place $\overline{I}_i$ in $\hat{I}_{i,2}$. This means $\hat{I}_{i,j}, j \in \{1, 3, 4\}$ are left blank.

We adapt a convolution-deconvolution structure for our scan completion network, denoted $g_\phi$. As shown in Figure 2, we use separate layers to extract information from color, depth, and normal input, and concatenate the resulting feature maps. Note that we stack the source and transformed target scan in each of the color, normal, depth components to provide the network more information. Only the source scan is shown for simplicity. Since designing the completion network is not the major focus of this paper, we leave the technical details to supplementary material.

### 3.4. Relative Pose Module

We proceed to describe the proposed relative pose module denoted as $h_\gamma(\overline{S}_1, \overline{S}_2) \to (R, t)$. This module essentially fits $(R, t)$ to consistent correspondences established between points sampled from $\overline{S}_1$ and $\overline{S}_2$. To this end, we first compute a point set $\mathcal{Q}_i$ from each $\overline{S}_i$. Due to noisy scan completions, we define $\mathcal{Q}_i$ by combining SIFT keypoints [28] in the observed region of $\overline{S}_i$ and uniform sample points in the remaining region. Let $\hat{\mathcal{C}} \subset \mathcal{Q}_1 \times \mathcal{Q}_2$ denote all correspondences between SIFT feature points in one scan and all points in the other scan. Our goal is to simultaneously extract a subset of correspondences from $\mathcal{C} \subset \hat{\mathcal{C}}$ and fit $(R, t)$ to these selected correspondences. For efficiency, we remove a correspondence $c = (q_1, q_2)$ from $\hat{\mathcal{C}}$ whenever $\exp(-\|\boldsymbol{f}(q_1) - \boldsymbol{f}(q_2)\|^2/2/\gamma_1^2) \leq 10^{-2}$.

The technical challenge of extracting correct correspondences is that due to imperfect scan completions, many correspondences with similar descriptors are still outliers. We address this challenge by combining spectral matching [24] and robust fitting [2]. Specifically, let $x_c \in \{0, 1\}, \forall c \in \hat{\mathcal{C}}$ be the indicator of $c$. We compute $(R, t)$ by solving

$$\underset{\{x_c\}, R, t}{\text{maximize}} \sum_{c, c' \in \mathcal{C}} w_\gamma(c, c') x_c x_{c'} \big(\delta - r_{(R,t)}(c) - r_{(R,t)}(c')\big)$$
$$\text{subject to } \sum_{c \in \mathcal{C}} x_c^2 = 1 \qquad (1)$$

As we will define next, $w(c, c')$ is a consistency score of the correspondence pair $(c, c')$, and $r_{(R,t)}(c)$ is a robust regression loss between $(R, t)$ and $c$. Note that the same as [24], (1) relaxes the constraint that each $x_c$ is binary to $\sum_{c \in \mathcal{C}} x_c^2 = 1$. $\delta$ is set to be 50 pixels in our experiments. Intuitively, (1) seeks to extract a subset of correspondences that have large pairwise consistency scores and can be fit well by $(R, t)$.

We define $w(c, c')$, where $c = (q_1, q_2)$ and $c' = (q_1', q_2')$, by combining five consistency measures. The first one mea-
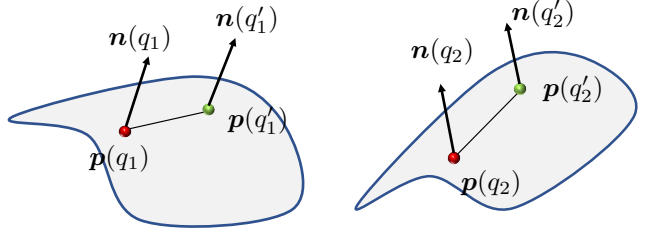


Figure 3: The geometry consistency constraints are based on the fact that rigid transforms preserve lengths and angles.

sures consistency in descriptors:

$$\Delta_1^2(c, c') := \|\boldsymbol{f}(q_1) - \boldsymbol{f}(q_2)\|^2 + \|\boldsymbol{f}(q_1') - \boldsymbol{d}(q_2')\|^2. \quad (2)$$

The remaining four terms measure geometric consistency in edge length and angles [37, 17] (See Figure 3):

$$\Delta_2(c, c') := \|\boldsymbol{p}(q_1) - \boldsymbol{p}(q_1')\| - \|\boldsymbol{p}(q_2) - \boldsymbol{p}(q_2')\|$$
$$\Delta_3(c, c') := \angle(\boldsymbol{n}(q_1), \boldsymbol{n}(q_1')) - \angle(\boldsymbol{n}(q_2), \boldsymbol{n}(q_2'))$$
$$\Delta_4(c, c') := \angle(\boldsymbol{n}(q_1), \boldsymbol{p}(q_1)\boldsymbol{p}(q_1')) - \angle(\boldsymbol{n}(q_2), \boldsymbol{p}(q_2)\boldsymbol{p}(q_2'))$$
$$\Delta_5(c, c') := \angle(\boldsymbol{n}(q_1'), \boldsymbol{p}(q_1)\boldsymbol{p}(q_1')) - \angle(\boldsymbol{n}(q_2'), \boldsymbol{p}(q_2)\boldsymbol{p}(q_2'))$$

We now define

$$w_\gamma(c, c') = \exp\Big(-\frac{1}{2}\sum_{i=1}^{5}\big(\frac{\Delta_i(c, c')}{\gamma_i}\big)^2\Big) \qquad (3)$$

where $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ are hyper-parameters associated with the consistency measures.

We define the robust rigid regression loss in (1) as

$$r_{(R,t)}(c) = \big(\|R\boldsymbol{p}(q_1) + t - \boldsymbol{p}(q_2)\|^2 + \|R\boldsymbol{n}(q_1) - \boldsymbol{n}(q_2)\|^2\big),$$

We perform alternating maximization to optimize (1). When $R$ and $t$ are fixed, (1) reduces to

$$\max_{x_c} \sum_{c, c'} a_{cc'} x_c x_{c'} \quad \text{subject to} \sum_c x_c^2 = 1, \qquad (4)$$

where $a_{cc'} := w_\gamma(c, c')\big(\delta - r_{(R,t)}(c) - r_{(R,t)}(c')\big)$. It is clear that the optimal solution $\{x_c\}$ is given by the maximum eigenvector of $A = (a_{cc'})$ (c.f. [33]). Likewise, when $\{x_c\}$ is fixed, (1) reduces to

$$\min_{R, t} \sum_{c \in \mathcal{C}} a_c r_{(R,t)}(c), \quad a_c := x_c \sum_{c' \in \mathcal{C}} w_\gamma(c, c') x_{c'}. \quad (5)$$

We solve (5) using iterative reweighted least squares (or IRLS). The step exactly follows [2] and is left to Section A.2 of the supp. material. In this paper, we use 5 iterations between spectral matching and robust fitting.

Our approach essentially combines the strengths of IRLS and spectral matching. IRLS is known to be sensitive to large outlier ratios (c.f. [9]). In our formulation, this limitation is addressed by spectral matching, which detects the strongest consistent correspondence subset. On the other hand, spectral matching, which is a relaxation of a binary-integer program, does not offer a clean separation between inliers and outliers. This issue is addressed by using IRLS.

## 3.5. Network Training

We train the proposed network by utilizing training data of the form $\mathcal{P}_{\text{train}} = \{(\{(\overline{I}_i, \overline{S}_i^{\star})\}, T^{\star})\}$, where each instance collects two input scans, their corresponding completions, and their relative pose. Network training proceeds in two phases. The first phase learns each individual module, and the second phase performs fine-tuning.

### 3.5.1 Learning Each Individual Module

**Learning semantic descriptors.** Since color, depth, normals, and semantic labels are all pre-specified, we only learn the semantic descriptor channels $\boldsymbol{f}_\theta$ introduced in Section 3.2. To this end, we first define a contrastive loss on the representation of scan completions for training globally discriminative descriptors:

$$\mathcal{L}_{des}(\overline{S}_1, \overline{S}_2) := \sum_{(q_1, q_2) \in \mathcal{G}(S_1, S_2)} \|\boldsymbol{f}(q_1) - \boldsymbol{f}(q_2)\|$$
$$+ \sum_{(q_1, q_2) \in \mathcal{N}(S_1, S_2)} \max(0, D - \|\boldsymbol{f}(q_1) - \boldsymbol{f}(q_2)\|), \quad (6)$$

where $\mathcal{G}(S_1, S_2)$ and $\mathcal{N}(S_1, S_2)$ collect randomly sampled corresponding point pairs and non-corresponding point pairs between $S_1$ and $S_2$, respectively. $D$ is set to 0.5 in our experiments. We then solve the following optimization problem to learn semantic descriptors:

$$\min_\theta \sum_{(\{(\overline{I}_i, \overline{S}_i^{\star})\}, T^{\star}) \in \mathcal{P}_{\text{train}}} \mathcal{L}_{des}(\overline{S}_1, \overline{S}_2) \quad s.t. \quad \boldsymbol{f} = \boldsymbol{f}_\theta \quad (7)$$

In our experiments, we train 100k iterations with batch size 2 using ADAM [23].

**Learning completion modules.** We train the completion network $g_\phi$ by combining a regression loss and a contrastive descriptor loss:

$$\min_\phi \sum_{(\{(\overline{I}_i, \overline{S}_i^{\star})\}, T^{\star}) \in \mathcal{P}_{\text{train}}} \mathop{E}_{T \sim \mathcal{N}(T^{\star}, \Sigma)} \Big( \|g_\phi(\hat{I}(\overline{I}_1, \overline{I}_2, T)) - S_1\|_{\mathcal{F}}^2$$
$$+ \lambda \mathcal{L}_{des}(\overline{S}_1, g_\phi(\hat{I}(\overline{I}_1, \overline{I}_2, T))) \Big)$$

where we set $\lambda = 0.01$. $\hat{I}(\overline{I}_1, \overline{I}_2, T)$ denotes the concatenated input of $\overline{I}_1$ and transformed $\overline{I}_2$ using $T$. Again, we train 100k iterations with batch size 2 using ADAM [23].

The motivation of the contrastive descriptor loss is that the completion network does not fit the training data perfectly, and adding this term improves the performance of descriptor matching. Also note that the input relative pose is not perfect during the execution of the entire network; thus we randomly perturb the relative pose in the neighborhood of each ground-truth for training.

**Pre-training relative pose module.** We pre-train the relative pose module using the results of the scan completion module:

$$\min_\gamma \sum_{(\{(\overline{I}_i, \overline{S}_i^{\star})\}, T^{\star}) \in \mathcal{P}_{\text{train}}} \|h_\gamma(\overline{S}_1, \overline{S}_2) - T^{\star}\|_{\mathcal{F}}^2. \quad (8)$$

For optimization, we employ finite-difference gradient descent with backtracking line search [31], which only requires evaluating the values of the objective function with respect to different hyper-parameters. In our experiments, the training converges in 30 iterations.

### 3.5.2 Fine-tuning Relative Pose Module

Given the pre-trained individual modules, we could fine-tune the entire network together. However, we find that training is hard to converge and the test accuracy even drops. Instead, a more effective fine-tuning strategy is to just optimize the relative pose modules. In particular, we allow them to have different hyper-parameters to accommodate specific distributions of the completion results at different alternating iterations. Specifically, let $\gamma$ and $\gamma_t$ be the hyper-parameters of the first pairwise matching module and the pairwise matching module at iteration $t$, respectively. With $T^{t_{\max}}(\overline{I}_1, \overline{I}_2)$ we denote the output of the entire network. We solve the following optimization problem for fine-tuning:

$$\min_{\gamma, \{\gamma_t\}} \sum_{(\{(\overline{I}_i, \overline{S}_i^{\star})\}, T^{\star}) \in \mathcal{P}_{\text{train}}} \|T^{t_{\max}}(\overline{I}_1, \overline{I}_2) - T^{\star}\|_{\mathcal{F}}^2. \quad (9)$$

Similar to (8), we again employ finite-difference gradient descent with backtracking line search [31] for optimization. In our experiments, the training converges in 20 iterations.

## 4. Experimental Results

In this section, we present an experimental evaluation of the proposed approach. We begin with describing the experimental setup in Section 4.1. We then present an analysis of our results in Section 4.2. Finally, we present an ablation study in Section 4.3.

### 4.1. Experimental Setup

#### 4.1.1 Datasets

We perform experimental evaluation on three datasets. **SUNCG** [39] is a synthetic dataset that collects 45k different 3D scenes, where we take 9892 bedrooms for experiments. For each room, we sample 25 camera locations around the room center. The field of view is set as $90°$ both horizontally and vertically. From each camera pose, we collect an input scan and the underlying ground-truth completion stored in its local coordinate system. We allocate 80% of rooms for training and the rest for testing. **Matterport** [3] is a real dataset that collects 925 different 3D scenes. Each room was reconstructed from a real indoor

| | SUNCG | | | | | | | | Matterport | | | | | | | | ScanNet | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rotation | | | | Trans. | | | | Rotation | | | | Trans. | | | | Rotation | | | | Trans. | | | |
| | 3° | 10° | 45° | Mean | 0.1 | 0.25 | 0.5 | Mean | 3° | 10° | 45° | Mean | 0.1 | 0.25 | 0.5 | Mean | 3° | 10° | 45° | Mean | 0.1 | 0.25 | 0.5 | Mean |
| 4PCS([0.5,1]) | 64.3 | 83.7 | 87.6 | 21.0 | 68.2 | 74.4 | 79.0 | 0.30 | 42.7 | 65.7 | 80.3 | 33.4 | 52.6 | 64.3 | 69.0 | 0.46 | 25.3 | 48.7 | 80.1 | 31.2 | 36.9 | 43.2 | 59.8 | 0.52 |
| GReg([0.5,1]) | 85.9 | 91.9 | 94.1 | 10.3 | 86.9 | 89.3 | 90.7 | 0.16 | 80.8 | 89.2 | 92.1 | 12.0 | 84.8 | 88.5 | 90.6 | 0.17 | 58.9 | 84.4 | 88.8 | 16.3 | 81.7 | 85.8 | 88.6 | **0.19** |
| CGReg([0.5,1]) | 90.8 | 92.9 | 93.9 | 9.8 | 87.3 | 90.7 | 92.8 | 0.13 | 90.3 | 90.8 | 93.1 | 10.1 | 89.4 | 89.6 | 91.6 | 0.14 | 59.0 | 75.7 | 88.1 | 18.0 | 62.1 | 77.7 | 86.9 | 0.23 |
| DL([0.5, 1]) | 0.0 | 0.0 | 15.9 | 81.4 | 0.0 | 1.9 | 8.5 | 1.60 | 0.0 | 0.0 | 9.9 | 83.8 | 0.0 | 3.3 | 6.6 | 1.77 | 0.0 | 0.0 | 30.0 | 61.3 | 0.0 | 0.1 | 0.7 | 3.31 |
| Ours-nc.([0.5,1]) | 88.6 | 94.7 | 97.6 | 4.3 | 83.4 | 92.6 | 95.9 | **0.10** | 90.5 | 97.6 | 98.9 | 2.3 | 93.7 | 96.9 | 98.9 | **0.04** | 57.2 | 80.6 | 90.5 | 13.9 | 66.3 | 79.6 | 85.9 | 0.24 |
| Ours-nr.([0.5,1]) | 90.0 | 96.0 | 97.8 | 4.3 | 83.8 | 94.4 | 96.5 | **0.10** | 85.9 | 97.7 | 99.0 | 2.7 | 88.9 | 94.6 | 97.2 | 0.07 | 51.0 | 78.3 | 91.2 | **12.7** | 63.7 | 79.2 | 86.8 | 0.22 |
| Ours([0.5, 1]) | 90.9 | 95.9 | 97.8 | **4.0** | 83.6 | 94.3 | 96.6 | **0.10** | 89.5 | 98.5 | 99.3 | **1.9** | 93.1 | 96.7 | 98.5 | 0.05 | 52.9 | 79.1 | 91.3 | **12.7** | 64.7 | 78.6 | 86.0 | 0.23 |
| 4PCS([0.1,0.5]) | 4.9 | 10.6 | 13.7 | 113.0 | 4.0 | 5.3 | 7.1 | 1.99 | 4.2 | 16.2 | 25.9 | 87.0 | 5.0 | 8.1 | 10.0 | 2.19 | 1.5 | 7.1 | 30.0 | 82.2 | 2.5 | 3.1 | 3.1 | 1.63 |
| GReg([0.1,0.5]) | 35.1 | 45.4 | 50.3 | 64.1 | 35.8 | 40.3 | 43.6 | 1.29 | 19.2 | 26.8 | 34.9 | 73.8 | 24.2 | 27.2 | 28.4 | 1.68 | 11.4 | 25.0 | 33.3 | 86.5 | 18.1 | 21.7 | 23.4 | 1.31 |
| CGReg([0.1,0.5]) | 46.4 | 48.5 | 51.0 | 63.4 | 40.2 | 42.7 | 46.0 | 1.34 | 28.5 | 29.3 | 35.9 | 73.9 | 28.1 | 28.3 | 29.5 | 1.99 | 11.8 | 20.0 | 32.9 | 88.2 | 11.6 | 16.0 | 21.0 | 1.36 |
| DL([0.1, 0.5)) | 0.0 | 0.0 | 8.0 | 94.0 | 0.0 | 1.8 | 4.0 | 2.06 | 0.0 | 0.0 | 8.5 | 94.3 | 0.0 | 0.4 | 2.7 | 2.25 | 0.0 | 0.0 | 7.5 | 92.1 | 0.0 | 0.0 | 0.0 | 4.03 |
| Ours-nc.([0.1,0.5]) | 47.5 | 62.6 | 71.4 | 32.8 | 36.3 | 54.6 | 63.4 | 0.89 | 54.4 | 75.7 | 83.7 | 22.8 | 53.3 | 65.3 | 73.7 | 0.55 | 14.1 | 37.1 | 56.0 | 55.3 | 18.8 | 31.2 | 41.3 | 0.98 |
| Ours-nr.([0.1,0.5]) | 60.3 | 80.3 | 83.7 | 20.8 | 41.2 | 70.0 | 80.6 | 0.56 | 47.3 | 72.9 | 82.4 | 24.6 | 44.4 | 65.1 | 73.9 | 0.57 | 12.2 | 36.0 | 65.3 | 45.2 | 18.1 | 33.6 | 47.0 | 0.90 |
| Ours([0.1,0.5)) | 67.2 | 84.1 | 86.4 | **18.1** | 44.8 | 73.8 | 83.9 | **0.49** | 53.7 | 80.7 | 87.9 | **17.2** | 52.0 | 71.2 | 81.4 | **0.45** | 14.4 | 39.1 | 66.8 | **43.9** | 19.6 | 35.5 | 48.4 | **0.87** |
| DL([0.0, 0.1)) | 0.0 | 0.0 | 2.1 | 115.4 | 0.0 | 1.4 | 4.3 | 2.23 | 0.0 | 0.0 | 2.1 | 125.9 | 0.0 | 0.2 | 2.1 | 2.83 | 0.0 | 0.0 | 0.0 | 130.4 | 0.0 | 0.0 | 0.0 | 5.37 |
| Ours-nc.([0.0,0.1]) | 2.2 | 5.8 | 13.8 | 102.1 | 0.1 | 0.7 | 5.6 | 2.21 | 1.3 | 4.9 | 11.7 | 117.3 | 0.2 | 0.2 | 0.9 | 3.10 | 0.5 | 4.8 | 16.3 | 99.4 | 0.0 | 0.5 | 2.2 | 1.92 |
| Ours-nr.([0.0,0.1]) | 12.6 | 27.1 | 33.8 | 83.4 | 3.2 | 15.7 | 28.8 | 1.78 | 1.6 | 11.4 | 27.3 | 92.6 | 0.2 | 2.2 | 7.3 | 2.33 | 0.7 | 7.7 | 29.1 | 83.4 | 0.2 | 1.7 | 7.6 | 1.70 |
| Ours([0.0,0.1)) | 15.7 | 32.4 | 37.7 | **79.5** | 4.5 | 21.3 | 34.3 | **1.66** | 2.5 | 16.3 | 31.3 | **87.3** | 0.3 | 3.0 | 11.7 | **2.19** | 0.9 | 8.8 | 32.8 | **78.9** | 0.4 | 2.3 | 8.7 | **1.62** |

Table 1: Benchmark evaluation on our approach and baseline approaches. Ours-nc and Ours-nr stand for our method with the completion module and recurrent module removed, respectively. For the rotation component, we show the percentage of pairs whose angular deviations fall within $3°, 10°$, and $45°$, respectively. For the translation component, we show the percentage of pairs whose translation deviations fall within $0.1m, 0.25m, 0.5m$. We also show the mean errors. In addition, we show statistics for pairs of scans whose overlapping ratios fall into three intervals, namely, [50%, 100%], [10%, 50%], and [0%, 10%]. Average numbers are reported for 10 repeated runs on test sets.

room. We use their default train/test split. For each room, we pick 50 camera poses. The sampling strategy and camera configuration are the same as SUNCG. **ScanNet** [8] is a real dataset that collects 1513 rooms. Each room was reconstructed using thousands of depth scans from Kinect. For each room, we select every 25 frames in the recording sequence. For each camera location, we render the cube-map representation using the reconstructed 3D model. Note that unlike SUNCG and Matterport, where the reconstruction is complete, ScanNet's reconstruction is partial, i.e., there are much more areas in our cube-map representation that have missing values due to the incompleteness of ground truth. For testing, we sample 1000 pairs of scans (source and target scan are from the same room) for all datasets.

### 4.1.2  Baseline Comparison

We consider four baseline approaches:
**Super4PCS [30]** is a state-of-the-art non-deep learning technique for relative pose estimation between two 3D point clouds. It relies on using geometric constraints to vote for consistent feature correspondences. We use the author's code for comparison.
**Global registration (or GReg) [47]** is another state-of-the-art non-deep learning technique for relative pose estimation. It combines cutting-edge feature extraction and reweighted least squares for rigid pose registration. GReg is a more robust version than fast global registration (or FGReg) [46], which focuses on efficiency. We use the Open3D implementation of GReg for comparison.
**Colored Point-cloud Registration (or CGReg) [32]** is a combination of GReg and colored point-cloud registration,

where color information is used to boost the accuracy of feature matching. We use the Open3D implementation.
**Deep learning baseline (or DL)[29]** is the most relevant deep learning approach for estimating the relative pose between a pair of scans. It uses a Siamese network to extract features from both scans and regress the quaternion and translation vectors. We use the authors' code and modify their network to take in color, depth, and normal as input.

### 4.1.3  Evaluation Protocol

We evaluate the rotation component $R$ and translation component $t$ of a relative pose $T = (R, t)$ separately. Let $R^\star$ be the ground-truth, we report the relative rotation angle $acos(\frac{\|R^\star R^T\|_\mathcal{F}}{\sqrt{2}})$. Let $t^\star$ be the ground-truth translation. We evaluate the accuracy of $t$ by measuring $\|t - t^\star + (R - R^\star)c_{I_s}\|$, where $c_{I_s}$ is the barycenter of $I_s$.

To understand the behavior of each approach on different types of scan pairs, we divide the scan pairs into three categories. For this purpose, we first define the overlap ratio between a pair of scans $I_s$ and $I_t$ as $o(I_s, I_t) = |I_s \cap I_t| / \min(|I_s|, |I_t|)$. We say a testing pair $(I_s, I_t)$ falls into the category of significant overlap, small overlap, and non-overlap if $o(I_s, I_t) \geq 0.5$, $0.5 \geq o(I_s, I_t) \geq 0.1$, and $o(I_s, I_t) \leq 0.1$, respectively.

### 4.2. Analysis of Results

Table 1 and Figure 4 provide quantitative and qualitative results of our approach and baseline approaches. Overall, our approach outputs accurate relative pose estimations. In the following, we provide a detailed analysis under each
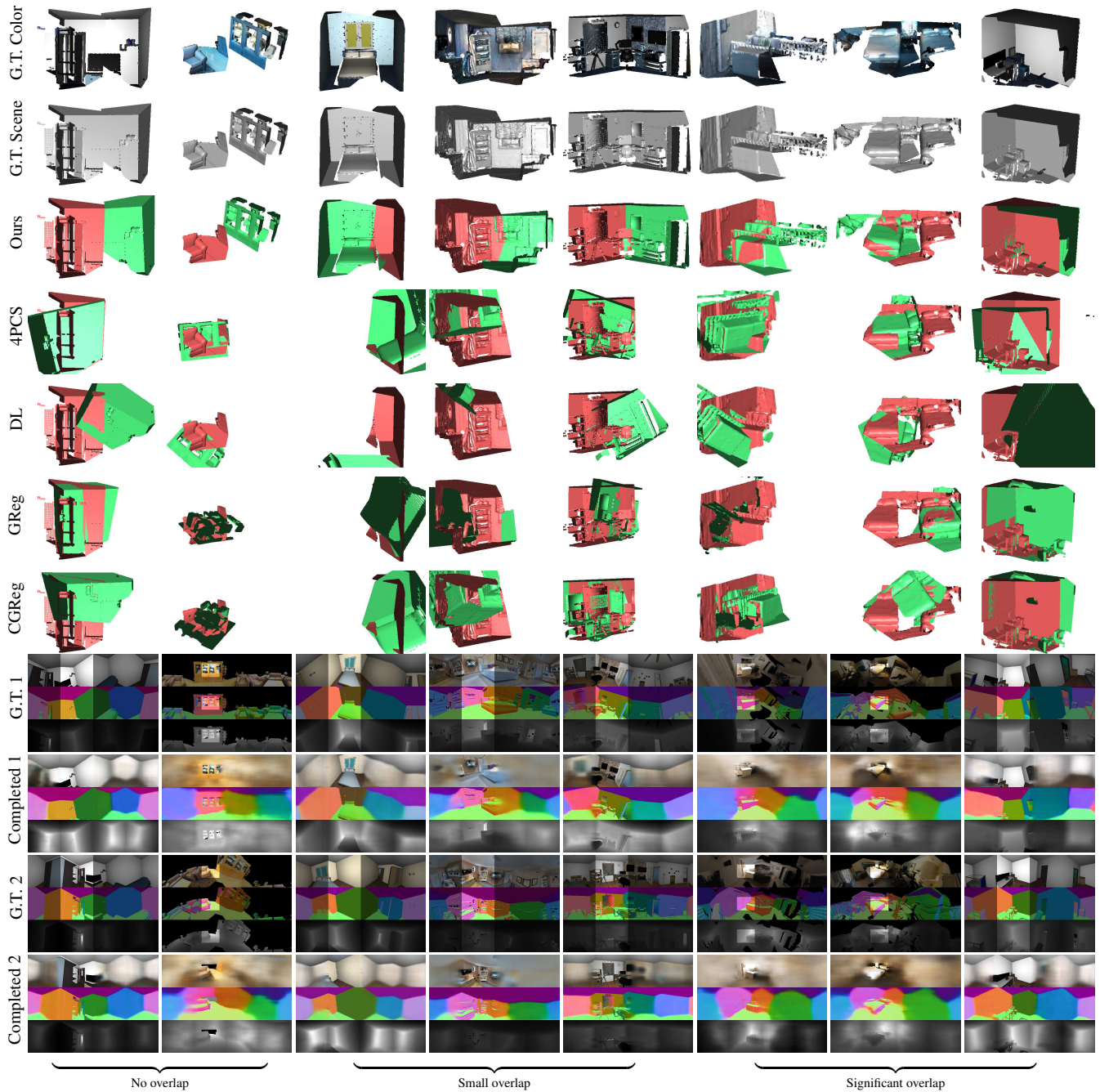
Figure 4: Qualitative results of our approach and baseline approaches. We show examples for the cases of no, small, and significant overlap. From top to bottom: ground-truth color and scene geometry, our pose estimation results (two input scans in red and green), baseline results (4PCS, DL, GReg and CGReg), ground-truth scene RGBDN and completed scene RGBDN for two input scans. The unobserved regions are dimmed. See Section 4.2 for details.

category of scan pairs as well as the scan completion results:

**Significant overlap.** Our approach outputs accurate relative poses in the presence of significant overlap. The mean error in rotation/translation of our approach is $3.9°/0.10m$, $1.8°/0.05m$, and $13.0°/0.23m$ on SUNCG, Matterport, and ScanNet, respectively, In contrast, the mean error in rotation/translation of the top performing methods only

achieves $9.8°/0.13m$, $10.1°/0.14m$, and $16.3°/0.19m$, respectively. Meanwhile, the performance of our method drops when the completion component is removed. This means that although there are rich features to match between significantly overlapping scans, performing scan completion still matters. Moreover, our approach achieves better relative performance on SUNCG and Matterport, as
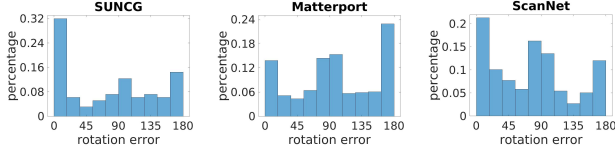
Figure 5: Error distribution of rotation errors of our approach on non-overlapping scans. See Section 4.2 for discussion.
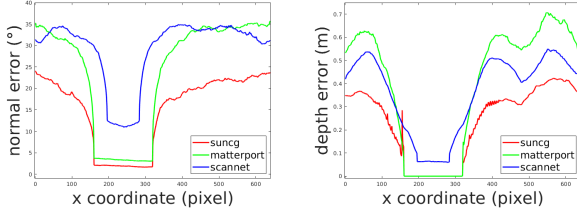


Figure 6: Mean errors in predicted normal and depth w.r.t the horizontal image coordinate. See Section 4.2 for discussion.

their field-of-views are wider than that of ScanNet.

**Small overlap.** Our approach outputs good relative poses in the presence of small overlap. The mean errors in rotation/translation of our approach are $20.1°/0.52m$, $16.3°/0.45m$, and $47.4°/0.90m$ on SUNCG, Matterport, and ScanNet, respectively. In contrast, the top-performing method only achieves mean errors $63.4°/1.29m$, $73.8°/1.68m$, and $82.2°/1.31m$, leaving a big margin from our approach. Moreover, the relative improvements are more salient than for scan pairs that possess significant overlap. This is expected as there are less points to match from the original scans, and scan completion provides more points to match.

**No overlap.** Our approach delivers encouraging relative pose estimations on the extreme non-overlapping scans. For example, in the first column of Figure 4, a television is separated into two parts in the source and target scans. Our method correctly assembles the two scans to form a complete scene. In the second example, our method correctly predicts the relative position of the sofa and bookshelf.

The mean rotation errors of our approach for these non-overlapping scans are $79.9°$, $87.9°$, and $81.8°$ on SUNCG, Matterport, and ScanNet, respectively. In contrast, the error of a random rotation is around $126.3°$. To further understand our approach, Figure 5 plots the error distribution of rotations. We can see a significant portion of the errors concentrate at $90°$ and $180°$, which can be understood from the perspective that our approach mixes different walls when performing pairwise matching. This is an expected behavior as many indoor rooms are symmetric.

**Scan-completion results.** Figure 6 plots the error distributions of predicted depth and normals with respect to the horizontal image coordinate. None that in our experiment the $[160, 320]$ region is observed for SUNCG/Matterport, and $[196, 284]$ for ScanNet. We can see that the errors are highly correlated with the distances to observed region, i.e.,

they are small in adjacent regions, and become less accurate when the distances become large. This explains why our approach leads to a significant boost on scan pairs with small overlaps, i.e., corresponding points are within adjacent regions.

### 4.3. Ablation Study

We consider two experiments to evaluate the effectiveness of the proposed network design. Each experiment removes one functional unit in the proposed network design. **No completion.** The first ablation experiment simply applies our relative pose estimation module on the input scans directly, i.e., without scan completions. The performance of our approach drops even on largely overlapping scans.

This means that it is important to perform scan completions even for partially overlapping scans. Moreover, without completion, our relative pose estimation module still possesses noticeable performance gains against the top-performing baseline GReg [47] on overlapping scans. Such improvements mainly come from combing spectral matching and robust fitting. Please refer to Section B of the supplementary material for in-depth comparison.

**No recurrent module.** The second ablation experiment removes the recurrent module in our network design. This reduced network essentially performs scan completion from each input scan and then estimates the relative poses between the scan completions. We can see that the performance drops in almost all the configurations. This shows the importance of the recurrent module, which leverages bi-scan completions to gradually improve the relative pose estimations.

### 5. Conclusions

We introduced an approach for relative pose estimation between a pair of RGB-D scans of the same indoor environment. The key idea of our approach is to perform scan completion to obtain the underlying geometry, from which we then compute the relative pose. Experimental results demonstrated the usefulness of our approach both in terms of its absolute performance when compared to existing approaches and the effectiveness of each module of our approach. In particular, our approach delivers encouraging relative pose estimations between extreme non-overlapping scans.

# References

[1] Dror Aiger, Niloy J. Mitra, and Daniel Cohen-Or. 4pointss congruent sets for robust pairwise surface registration. *ACM Trans. Graph.*, 27(3):85:1–85:10, Aug. 2008. 1, 2

[2] Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. Sparse iterative closest point. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, SGP '13, pages 113–123, Aire-la-Ville, Switzerland, Switzerland, 2013. Eurographics Association. 2, 4

[3] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676, 2017. 2, 5

[4] Qifeng Chen and Vladlen Koltun. Robust nonrigid registration by convex optimization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2039–2047, Washington, DC, USA, 2015. IEEE Computer Society. 2

[5] Taeg Sang Cho, Shai Avidan, and William T. Freeman. The patch transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1

[6] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 628–644, 2016. 3

[7] James M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 941–, Washington, DC, USA, 1999. IEEE Computer Society. 1

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443, 2017. 2, 6

[9] Ingrid Daubechies, Ronald A. DeVore, Massimo Fornasier, and C. Sinan Güntürk. Iteratively re-weighted least squares minimization: Proof of faster than linear rate for sparse recovery. In *42nd Annual Conference on Information Sciences and Systems, CISS 2008, Princeton, NJ, USA, 19-21 March 2008*, pages 26–29, 2008. 4

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2758–2766, 2015. 2

[11] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3-d rigid body transformations: A comparison of four major algorithms. *Mach. Vision Appl.*, 9(5-6):272–290, Mar. 1997. 2

[12] Natasha Gelfand, Niloy J. Mitra, Leonidas J. Guibas, and Helmut Pottmann. Robust global registration. In *Proceedings of the Third Eurographics Symposium on Geometry Processing*, SGP '05, Aire-la-Ville, Switzerland, Switzerland, 2005. Eurographics Association. 1, 2

[13] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *Int. J. Rob. Res.*, 31(5):647–663, Apr. 2012. 1

[14] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. 2, 11, 12

[15] Qi-Xing Huang and Dragomir Anguelov. High quality pose estimation by aligning multiple scans to a latent map. In *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*, pages 1353–1360. IEEE, 2010. 3

[16] Qi-Xing Huang, Bart Adams, Martin Wicke, and Leonidas J. Guibas. Non-rigid registration under isometric deformations. In *Proceedings of the Symposium on Geometry Processing*, SGP '08, pages 1449–1457, Aire-la-Ville, Switzerland, Switzerland, 2008. Eurographics Association. 2

[17] Qi-Xing Huang, Simon Flöry, Natasha Gelfand, Michael Hofer, and Helmut Pottmann. Reassembling fractured objects by geometric matching. *ACM Trans. Graph.*, 25(3):569–578, July 2006. 1, 2, 4

[18] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925, 2016. 2

[19] Dinesh Jayaraman, Ruohan Gao, and Kristen Grauman. Shapecodes: Self-supervised feature learning by lifting views to viewgrids. In *ECCV*, 2018. 2

[20] D. Jayaraman and K. Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] Oliver van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel CohenOr. A Survey on Shape Correspondence. *Computer Graphics Forum*, 2011. 2

[22] Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NIPS*, page to appear, 2018. 2

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[24] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pages 1482–1489, Washington, DC, USA, 2005. IEEE Computer Society. 2, 4

[25] Xin Li and S. S. Iyengar. On computing mapping of 3d objects: A survey. *ACM Comput. Surv.*, 47(2):34:1–34:45, Dec. 2014. 2

[26] Chenxi Liu, Alexander G. Schwing, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Rent3d: Floor-plan priors for monocular layout estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3413–3421, 2015. 1

[27] Jonathan Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1601–1609, 2014. 2

[28] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 4

[29] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 675–687. Springer, 2017. 2, 6

[30] Nicolas Mellado, Dror Aiger, and Niloy J. Mitra. Super 4pcs fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33(5):205–215, 2014. 1, 2, 6

[31] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006. 5

[32] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 143–152, 2017. 2, 6

[33] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998. 4

[34] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544, 2016. 2, 3

[35] Charles Ruizhongtai Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5648–5656, 2016. 3

[36] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *European Conference on Computer Vision*, pages 292–309. Springer, 2018. 2, 12

[37] Y. Shan, B. Matei, H. S. Sawhney, R. Kumar, D. Huber, and M. Hebert. Linear model hashing and batch ransac for rapid and accurate object recognition. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pages 121–128, Washington, DC, USA, 2004. IEEE Computer Society. 4

[38] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006. 1

[39] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5

[40] Shuran Song, Andy Zeng, Angel X Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[41] Jrgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, pages 573–580. IEEE, 2012. 1

[42] Johan W. Tangelder and Remco C. Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl.*, 39(3):441–471, Sept. 2008. 2

[43] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, volume 5, page 6, 2017. 2

[44] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016. 3

[45] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1696–1704. Curran Associates, Inc., 2016. 3

[46] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 766–782, 2016. 1, 2, 6

[47] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *CoRR*, abs/1801.09847, 2018. 2, 6, 8

[48] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 2

[49] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with RGB-D cameras. *Comput. Graph. Forum*, 37(2):625–652, 2018. 1

[50] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2