Pose Synchronization under Multiple Pair-wise Relative Poses

Yifan Sun Qixing Huang
The University of Texas at Austin
2317 Speedway, Austin, Texas, 78712
yifansun12@utexas.edu huangqx@cs.utexas.edu

Abstract

Pose synchronization, which seeks to estimate cons
absolute poses among a collection of o
relative poses estimated between pair

applications. This paper studies an extreme setting where multiple relative pose estimates exist between each object pair, and the majority is incorrect. Popular methods that solve pose synchronization via recovering a low-rank matrix that encodes relative poses in block fail under this extreme setting. We introduce a three-step algorithm for pose synchronization under multiple relative pose inputs. The first step performs diffusion and clustering to compute the candidate poses of the input objects. We present a theoretical result to justify our diffusion formulation. The second step jointly optimizes the best pose for each object. The final step refines the output of the second step. Experimental results on benchmark datasets of structure-from-motion and scan-based geometry reconstruction show that our approach offers more accurate absolute poses than state-of-the-art pose synchronization techniques.

1. Introduction

Pose synchronization, which seeks to estimate absolute object poses from noisy relative poses estimated between object pairs, is a fundamental problem in many inverse applications in vision and graphics. Examples include multi-view structure from motion [38], 3D reconstruction from RGB-D scans [21], and reassembling fractured objects [14]. This problem has received great process during the past two decades, starting early greedy approaches [14, 21] to recent optimization-based approaches [2, 4, 11, 15, 17, 19, 20, 26–31, 34, 37]. However, existing approaches assume that there is only one relative pose for each object pair, and most relative poses are inliers. This assumption breaks when relative pose estimation is challenging, e.g., in 3D reconstruction from sparse views. The correct poses may differ from the top-ranked relative poses obtained by a pairwise matching method.

In this paper, we study a new pose synchronization

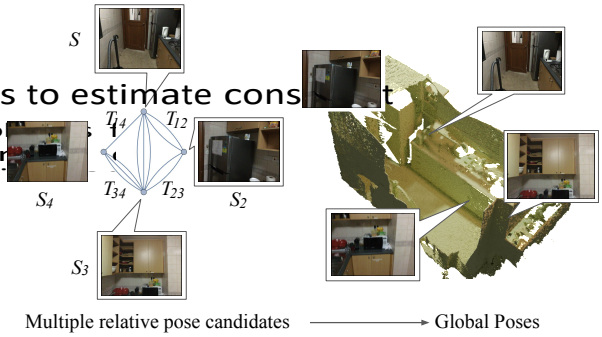

Figure 1. Our approach takes multiple candidate relative poses between pairs of objects as input and outputs absolute poses of the input objects for geometry reconstruction.

setting, where there are multiple relative pose estimates between an object pair, and most of them may be incorrect. This setting is quite popular, e.g., when overlapping ratios are low, or objects possess partial symmetries. Our approach proceeds in three simple steps. The first step computes for each object a set of candidate poses. This step is based on the fact that a correct relative pose between any object and a root object shall be realized by composing relative poses along multiple paths that connect them. We introduce an iterative procedure that alternates between diffusion and clustering to compute candidate poses. The second step solves a Markov Random Field (MRF) inference problem to jointly selects the best pose for each object so that the induced relative pose agrees with the input relative poses. The third step performs robust optimization to fine-tune the absolute poses of input objects.

Our approach's novelty is a diffusion formulation that synchronizes potentially multiple relative poses between object pairs into candidate poses for each object. The formulation, which utilizes a mixture model, is accurate, robust to noise, and theoretically justified. The resulting candidate poses enable a simple MRF approach via the projected power method [36]. Compared to prior MRF formulations [8,33] that are based on uniform sampling, our approach does not suffer from discretization errors.

We have evaluated our approach on benchmark datasets of multi-view structure-from-motion and geometry reconstruction from depth scans. Experimental results

show that our approach outperforms state-of-the-art pose synchronization approaches.

2. Related Works

Existing approaches fall into three categories. The first category of methods are based on computing a spanning tree among the input graph of relative poses [14, 21]. However, as there are potentially an exponential number of spanning trees in a graph, these approaches are either slow or easily give rise to sub-optimal solutions.

The second category of methods is based on the fact that consistent relative poses can be encoded as a low-rank matrix while the relative poses predicted using pairwise techniques are noisy measurements of the matrix blocks of this low-rank matrix, c.f. [15]. This leads to continuous optimization approaches for synchronization, which recover the underlying low-rank matrix from noisy measurements of its blocks. State-of-the-art techniques include semidefinite programming relaxation [11, 15, 16, 28, 30, 34, 37], spectral techniques [1–3, 17, 20, 23, 26, 27, 29, 31, 32], and non-convex optimization [4, 6, 19, 24, 25, 39]. However, the basis of low-rank matrix recovery is that the spectral gap of the matrix that encodes good input relative poses separates from the matrix that encodes outliers in input relative poses. This assumption is valid mainly when most input relative poses are correct. It does not apply when multiple input relative poses exist between a pair of objects.

The third category of methods solves an inference problem on Markov random field (or MRF) [8, 13, 18]. The key idea is to discretize the rigid pose of each object and then jointly pick the best poses of all objects so that the induced relative poses best align with the input relative poses. The advantage of this approach is that it can incorporate multiple input relative poses between pairs of objects. However, this approach's performance heavily depends on the discretization error when sampling the pose space of each object, which significantly affects the energy landscape of the resulting MRF inference problem. SFM-MRF [8] performs uniform sampling on rotations and translations, which leads to a large-scale MRF problem with many competing solutions that are difficult to solve. In contrast, our approach uses an iterative procedure to generate candidate poses from the input relative poses. This leads to a small-scale MRF problem. Moreover, the resulting MRF is easy to solve, allowing us to adopt an iterative algorithm with guarantees.

Our approach for generating candidate poses alternates between diffusing candidate poses to neighboring objects and then clustering each object's propagated poses from neighboring objects. This paradigm was first proposed in [18] for computing point-based maps among a collection of objects. A recent work [33] extended the idea for pose synchronization among symmetric objects. However, both formulations still use sampling for discretization, i.e., samples on surfaces for point-based maps and partitioning of the pose space into bins for clustering. In contrast, we introduce a continuous formulation that performs clustering without sampling the pose space. We also provide provable guarantees on the quality of the resulting candidate poses.

Several approaches [4, 8] have studied continuous optimization under robust norms to refine absolute poses from relative poses. A popular paradigm is to employ iteratively reweighted non-linear least squares (IRNLLS) for optimization. Thanks to a highly accurate initial solution for pose refinement, our approach uses the Gauss-Newton method for optimization, which outperforms IRNLLS.

3. Approach

We begin with the problem statement and an overview of the proposed three step approach in Section 3.1. Section 3.2 to Section 3.4 present the technical details of each step.

3.1. Problem Statement and Approach Overview

Problem statement. The input to our approach is given by an observation graph $G = (S, E)$, where $S = \{S_1, \cdots, S_n\}$ denotes the input objects, and where E denotes the edges. There may be multiple edges between one object pair, encoding multiple outputs of a pairwise matching algorithm. Let $E(i, j)$ denote the subset of edges from object S_i to S_j. Each edge $e \in E(i, j)$ is associated with a rigid transformation $T_e = (R_e, t_e) : \Sigma_i \rightarrow \Sigma_j$ that predicts a relative pose between the local coordinate system Σ_i of object S_i and the local coordinate system Σ_j of object S_j. We assume that a reverse edge $e' \in E(j, i)$ exists for all $e \in E(i, j)$ and the associated relative pose $T_{e'} = T_e^{-1}$ is the inverse of T_e. Our goal is to predict the absolute pose $T_i : \Sigma_i \rightarrow \Sigma$ of object S_i in a world coordinate system Σ.

Approach overview. As shown in Figure 2, our approach proceeds in three steps. The first step computes a candidate set of absolute poses $T_i \in SE(3)$ for each object S_i. Here the absolute poses are expressed with respect to the local coordinate system of the object that has the maximum degree in G. We present a principled approach that alternates between a pose propagation and a pose clustering. The second step joints picks the optimal absolute pose T_i^g of each object S_i by solving a quadratic relaxation of MRF inference. The third step performs continuous optimization to refine the absolute pose T_i of each object S_i.

3.2. Step I: Candidate Pose Generation

Without losing generality, we assume that the object with the maximum degree is S_1. The underlying principle for computing the candidate pose set T_i is that a candidate pose should be realized by composing relative poses along many paths in G that connects S_1 and S_i. Formally speaking, we first define pose composition along paths.

Definition 1 Consider a path $p = (S_{p_0} S_{p_1} \cdots S_{p_{l(p)}})$ that connects S_{p_0} and $S_{p_{l(p)}}$ where $l(p)$ is the number of edges of p. The composite relative pose $T_p = (R_p, t_p)$ along p is

$$T_p := T_{p_{l(p)-1} p_{l(p)}} \circ \cdots \circ T_{p_0 p_1}. \quad (1)$$

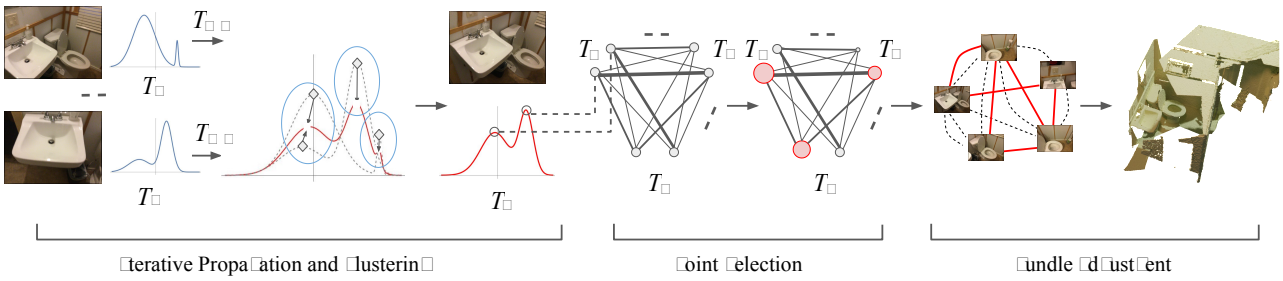

Figure 2. This figure illustrates our three-step approach. Given relative poses predicted between pairs of objects, the first step computes a candidate set of absolute poses for each object. This is done by alternating between pose propagation and pose clustering. The second step jointly selects the best absolute pose for each object. The third step performs bundle adjustment to refine object poses.

Next, we define the induced probability distributions from a rigid pose. Similar to Gaussian mixture models used for mean-shift clustering [7], they are used to aggregate composite relative poses along different paths.

Definition 2 The induced probability distributions over the space of rotations $SO(3)$ and the space of translations $\mathbb{R}^3$ from a rigid pose $T = (R, t)$ are given by

$$P^{\sigma_R,R}(R') = e^{-\frac{\|R'-R\|_F^2}{2\sigma_R^2}}, \quad R' \in SO(3) \tag{2}$$

$$P^{\sigma_t,t}(t') = e^{-\frac{\|t'-t\|^2}{2\sigma_t^2}}, \quad t' \in \mathbb{R}^3 \tag{3}$$

where σ_R and σ_t are hyper-parameters of P^R and P^t, respectively.

Now we are ready to define an induced probability distributions on rotations and translations of a pose graph.

Definition 3 Given a pose graph G, the un-normalized induced probability distributions on rotations and translations of object S_i in the local coordinate system of S_1 are defined as

$$Q^{\{\sigma_R^l\},R,S_i}_{\alpha,l_{max},l_0}(R') = \sum_{l=1}^{l_{max}} \frac{\alpha^l}{(l+l_0)!} \sum_{p \in P_{1i}^{(l)}} P^{\sigma_R^l,R_p}(R'), \tag{4}$$

$$Q^{\{\sigma_t^l\},t,S_i}_{\alpha,l_{max},l_0}(t') = \sum_{l=1}^{l_{max}} \frac{\alpha^l}{(l+l_0)!} \sum_{p \in P_{1i}^{(l)}} P^{\sigma_t^l,t_p}(t'), \tag{5}$$

where $P_{1i}^{(l)}$ collects all the paths from S_1 to S_i whose length is l; l_0 and l_{max} are the minimum and maximum lengths, respectively, of the paths we consider.

There are several hyper-parameters in (4) and (5). α is a decaying parameter that penalizes long paths; σ_R^l and σ_t^l are hyper-parameters of paths of length l. In this paper, we set $\sigma_R^l(\sigma_t^l) = \sqrt{l}\sigma_R(\sigma_t)$ where σ_R and σ_t are universal hyper-parameters (to be defined later) shared by all paths. The coefficient $\sqrt{l}$ is motivated from the fact the variance of composing l independent observations scales as $O(\sqrt{l})$.

In the following, we show that under mild conditions, $Q^{\{\sigma_R^l\},R,S_i}_{\alpha,l_{max},l_0}(R')$ and $Q^{\{\sigma_t^l\},t,S_i}_{\alpha,l_{max},l_0}(t')$ converge to $Q^{\{\sigma_R^l\},R,S_i}_{\alpha,\infty,0}(R')$ and $Q^{\{\sigma_t^l\},t,S_i}_{\alpha,\infty,0}(t')$, respectively.

Proposition 1 When $\alpha > 0$, then $Q^{\{\sigma_R^l\},R,S_i}_{S_i,\infty,0}$ and $Q^{\{\sigma_t^l\},t,S_i}_{S_i,\infty,0}$ are well-defined and

$$\|Q^{\{\sigma_R^l\},R,S_i}_{\alpha,l_{max},0} - Q^{\{\sigma_R^l\},R,S_i}_{\alpha,\infty,0}\|_1 \le \sqrt{\frac{e^{\alpha\lambda_1}}{l_{max}!}(\alpha\lambda_1)^{l_{max}}}n, \tag{6}$$

$$\|Q^{\{\sigma_t^l\},t,S_i}_{\alpha,l_{max},0} - Q^{\{\sigma_t^l\},t,S_i}_{\alpha,\infty,0}\|_1 \le \sqrt{\frac{e^{\alpha\lambda_1}}{l_{max}!}(\alpha\lambda_1)^{l_{max}}}n, \tag{7}$$

where λ_1 is spectral norm of the adjacency matrix of G.

To show the usefulness of the induced probability distributions, we next show that under a standard noise model, the global maximums of $Q^{\{\sigma_R^l\},R,S_i}_{\alpha,\infty,0}$ and $Q^{\{\sigma_t^l\},t,S_i}_{\alpha,\infty,0}$ recover the underlying ground-truth pose of S_i.

Proposition 2 (Informal) Suppose all T_{ij} are independent and identically follow a normal distribution with probability w_1 and a uniform distribution with probability w_2. Let P_1, P_2 be the PDFs of the two distributions. If $(1 - w_1^{\alpha\lambda_1})P_2(t) \le 1$, then almost surely the global maximums of $Q^{\{\sigma_R^l\},R,S_i}_{\alpha,\infty,0}$ and $Q^{\{\sigma_t^l\},t,S_i}_{\alpha,\infty,0}$ are with in the Δ-neighborhood of the ground-truth, where

$$\Delta = O(\lambda_1^{-\frac{1}{6}}(\ln \alpha\lambda_1)^{\frac{2}{3}}). \tag{8}$$

Moreover, this neighborhood is almost convex when we choose $\sigma_R, \sigma_t = O(\Delta)$.

Approximations of induced probabilities. Computing $Q^{\{\sigma_R^l\},R,S_i}_{\alpha,l_{max},l_0}$ and $Q^{\{\sigma_t^l\},t,S_i}_{\alpha,l_{max},l_0}$ directly is infeasible as the size of $P_{1i}^{(l)}$ quickly explodes when l increases. In the following, we present an iterative algorithm for approximating them. Specifically, let N_i denote the adjacent vertices of S_i in G. It is clear that the path set $P_{1i}^{(l)}$ admits the following recursion:

$$P_{1i}^{(l+1)} = \bigcup_{j \in N_i} \{(i,p) \mid p \in P_{1j}^{(l)}\}. \tag{9}$$

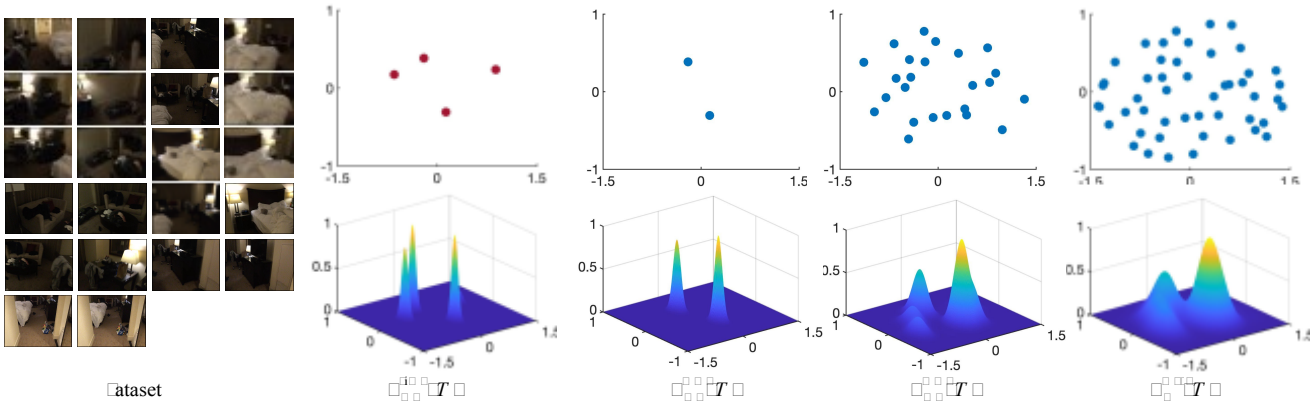

Figure 3. Candidate pose generation. (Col.1) A dataset with 22 scans with two pairwise matches per edge on average. (Col.2) Peak centers and the density function from our results using multi-dimensional scaling. (Col.3-5) Visualizations of $P_{12}^{(4)}(T)$, $P_{12}^{(8)}(T)$, and $P_{12}^{(12)}(T)$.

Applying (9), we obtain the following recursions:

$$Q_{\alpha,l,l'-1}^{\{\sigma_R^l\},R,S_i}(R') = \frac{\alpha}{l'}\sum_{(1,i)\in E} P^{\sigma_R^1,R_{1i}}(R') + \sum_{j\in N_i} Q_{\alpha,l-1,l'}^{\{\sigma_R^{l+1}\},R,S_j}(R_{ij}R') \quad (10)$$

$$Q_{\alpha,l,l'-1}^{\{\sigma_t^l\},t,S_i}(t') = \frac{\alpha}{l'}\sum_{(1,i)\in E} P^{\sigma_t^1,t_{1i}}(t') + \sum_{j\in N_i} Q_{\alpha,l-1,l'}^{\{\sigma_t^{l+1}\},t,S_j}(R_{ij}R' + t_{ij}) \quad (11)$$

It remains to control the complexity of $Q_{\alpha,l,l'}^{\{\sigma_R^l\},R,S_i}$ and $Q_{\alpha,l,l'}^{\{\sigma_t^l\},t,S_i}$ so that the number of mixture components in them do not explode. The key idea is to approximate each of them using a low complexity Gaussian mixture function (GMM) after applying (10) and (11). It is easy to see that both recursions (10) and (11) apply whenever induced probabilities are given by GMMs.

To simplify a GMM, we propose to detect its local maximums and fit a Gaussian function around each local minimum. Consider a GMM on $SE(3)$:

$$Q^T(T') = \sum_{k=1}^{K} w_k e^{-d^2_{\sigma_{k,R},\sigma_{k,t}}(T',T_k)}$$

$$d^2_{\sigma_{k,R},\sigma_{k,t}}(T',T_k) := \frac{\|R' - R_k\|_F^2}{2\sigma_{k,R}^2} + \frac{\|t' - t_k\|^2}{2\sigma_{k,t}^2} \quad (12)$$

where w_k and $(\sigma_{k,R},\sigma_{k,t})$ are the magnitude and variance parameters of the k-th component.

Proposition 3 $T' = (R', t')$ is a critical point of Q^T if

$$R' = UV^T,\ U\Sigma V^T = \sum_{k=1}^{K} \frac{w_k}{\sigma_{k,R}^2} e^{-d^2_{\sigma_{k,R},\sigma_{k,t}}(T',T_k)}$$

$$t' = \sum_{k=1}^{K} \frac{w_k}{\sigma_{k,t}^2} e^{-d^2_{\sigma_{k,R},\sigma_{k,t}}(T',T_k)} \quad (13)$$

where $U\Sigma V^T$ is the singular value decomposition (SVD).

Applying (13), we arrive at the following iterative procedure for computing local maximums of Q^T:

$$R^n \leftarrow UV^T,\ U\Sigma V^T = \sum_{k=1}^{K} \frac{w_k}{\sigma_{k,R}^2} e^{-d^2(T^c,T_k)} \cdot R_k,$$

$$t^n \leftarrow \sum_{k=1}^{K} \frac{w_k}{\sigma_{k,t}^2} e^{-d^2(T^c,T_k)} \cdot t_k \Big/ \sum_{k=1}^{K} \frac{w_k}{\sigma_{k,t}^2} e^{-d^2(T^c,T_k)}, \quad (14)$$

where $T^n (T^c)$ denote the pose at the next(current) iteration. The starting solutions for (14) are $T_k = (R_k, t_k)$.

After convergence, we detect distinctive local maximums. Decompose $\{1,\cdots,K\} = C_1 \cup \cdots C_{n_c}$ where each C_j collects initial poses T_k that converge to the same local maximum T_j. We then compute the optimal variance parameters and an optimal mixture weight associated with T_j via least square fitting.

Proposition 4 The optimal solution to

$$\min_{\sigma_R,\sigma_t,w} \int_{T'\in R^{3\times 4}} \Big| w e^{-d^2_{\sigma_R,\sigma_t}(T',T_j)} - \sum_{k\in C_j} w_k e^{-d^2_{\sigma_{k,R},\sigma_{k,t}}(T,T_k)} \Big|^2 \quad (15)$$

is given by

$$w_j = \sum_{k\in C_j} w_k \left(\frac{2\sigma_{k,R}^2}{\sigma_{j,R}^2 + \sigma_{k,R}^2}\right)^{\frac{9}{2}} \left(\frac{2\sigma_{k,t}^2}{\sigma_{j,t}^2 + \sigma_{k,t}^2}\right)^{\frac{3}{2}} \cdot e^{-\frac{\|R_j - R_k\|_F^2}{2(\sigma_{j,R}^2 + \sigma_{k,R}^2)}} e^{-\frac{\|t_j - t_k\|^2}{2(\sigma_{j,t}^2 + \sigma_{k,t}^2)}}$$

$$\sigma_{j,R}', \sigma_{j,t}' = \arg\max_{\sigma_R',\sigma_t'} \sum_{k\in C_j} w_k \left(\frac{2\sigma_R'^2\sigma_{k,R}^2}{\sigma_R'^2 + \sigma_{k,R}^2}\right)^{\frac{9}{2}} e^{-\frac{\|R_j - R_k\|_F^2}{2(\sigma_R'^2 + \sigma_{k,R}^2)}} \left(\frac{2\sigma_t'^2\sigma_{k,t}^2}{\sigma_t'^2 + \sigma_{k,t}^2}\right)^{\frac{3}{2}} e^{-\frac{\|t_j - t_k\|^2}{2(\sigma_t'^2 + \sigma_{k,t}^2)}} \quad (16)$$

We compute $(\sigma_{j,R}^{2}, \sigma_{j,t}^{2})$ in (16) via alternating maximization, starting from $\sigma_R' = \underset{k \in C_j}{\text{median}}(\sigma_{k,R})$ and $\sigma_t' = \underset{k \in C_j}{\text{median}}(\sigma_{k,t})$. See the supp. material for details.

After compressing the mixture components of rotations and translations, we keep the top-k ($k = 4$ in all of our experiments) components with maximum weights. Figure 3 illustrates the diffusion and clustering procedure on a dataset with 22 scans.

Hyper-parameters σ_R and σ_t. We compute them by detecting consistent cycles in the input graph. Specifically, we first use [12] to compute a collection of cycles C. For each cycle $c \in C$, we say it is consistent if the composite rotation R_c and translation t_c satisfy $\|R_c - I_3\| \leq \sqrt{|c|}\epsilon_R$ and $\|t_c\| \leq \sqrt{|c|}\epsilon_t$, where $|c|$ is the length of c. In all of our experiments, we set $\epsilon_R = 0.1$ and $\epsilon_t = 0.1\epsilon$ where ϵ is the average object size. Let C_{cons} denote all consistent cycles. We compute $\sigma_R = \text{median}_{c \in C_{cons}}(\|R_c - I_3\|/\sqrt{|c|})$ and $\sigma_t = \text{median}_{c \in C_{cons}}(\|t\|/\sqrt{|c|})$

3.3. Step II: Optimal Pose Selection

The second step takes the candidate poses of each object $T_i = \{T_{ix_i} = (R_{ix_i}, t_{ix_i}) | x_i \in [n_i]\}$ as input and jointly picks the optimal pose $T_i^g = (R_i^g, t_i^g) \in T_i$. We formulate this step as solving a second-order MRF which optimizes the following objective function:

$$\max_{x_1 \in [n_1], \cdots, x_n \in [n_{|S|}]} \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \sum_{e \in E(i,j)} \varphi_e(T_{ix_i}, T_{jx_j}) \quad (17)$$

where

$$\varphi_e(T_{ix_i}, T_{jx_j}) = \exp\left(-\frac{\|R_e^{in} R_{ix_i} - R_{x_j}\|_F^2}{2\sigma_R^2} - \frac{\|R_e^{in} t_{ix_i} + t_e^{in} - t_{jx_j}\|^2}{2\sigma_t^2}\right)$$

where $\|\cdot\|_F$ is the matrix Frobenius norm.

Since the candidate poses are computed via diffusion, they are more accurate than those obtained from sampling the space of rotations and translations (e.g., SFM-MRF [8]), which are subject to discretization errors. We find that (17) can be solved effectively using projected power method [36], which is a simple iterative procedure in contrast to more complex MRF inference techniques [8,22].

Specifically, denote $N = \sum_{i=1}^{n} n_i$. Let $A \in R^{N \times N}$ be a block matrix that stores the values $\varphi_e(T_{ix_i}, T_{jx_j})$, i.e., the elements of the (i,j)-th block A_{ij} are given by $A_{ij}(x_i, x_j) = \sum_{e \in E(i,j)} \varphi_e(T_{ik}, T_{jl}), x_i \in [n_i], x_j \in [n_j]$. Introduce the corresponding $n \times 1$ block indicator vector $v \in R^N$ whose element $v_{ix_i}, i \in [|S|], x_i \in [n_i]$ specifies the confidence on selecting T_{ix_i} as the pose for the $i-th$ object. Initialize $v_{ix_i}^{(0)} = \frac{1}{n_i}, x_i \in [n_i], i \in [|S|]$. At each iteration $t \geq 1$, we alternate between a matrix-vector multiplication operation and projection operation:

$$\hat{v}^{(t)} = Av^{(t-1)}, \quad v_i^{(t)} = P_{\Delta_{n_i}}(\mu_t \hat{v}_i). \quad (18)$$

where $P_{\Delta_{n_i}} : R^{n_i} \to \Delta_{n_i}$ is the projection operation on the simplex $\Delta_{n_i} := \{x | x \in R^{n_i}, 1^T x = 1, x \geq 0\}$:

$$P_{\Delta_{n_i}}(\hat{v}_i) := \underset{y \in \Delta_i}{\text{argmin}} \|\hat{v}_i - y\|^2.$$

$\mu_t \to \infty$ is a scaling parameter. We set $\mu_t = \|A\|_{1,\infty} t$ for all of our experiments. (18) is run $t_{max} = 200$ steps. After that, the selected pose is given by

$$T_i^g = T_{ix_i^*}, \quad x_i^* = \arg\max_{x_i} v_{ix_i}^{(t)}.$$

3.4. Step III: Pose Refinement

The third step performs continuous optimization to refine the object poses (R_i, t_i). This is achieved by using robust norms to extract inliers in the input pairwise rotations R_e^{in} and t_e^{in} translations for e in each set $E(i,j)$ that satisfy $R_e^{in} R_i = R_j$ and $R_e^{in} t_i + t_e^{in} = t_j$ approximately:

$$\min_{R_i} \sum_{e \in E(i,j)} \rho_R(\|R_e^{in} - R_j R_i^T\|_F) \quad \text{s.t. } R_1 = I_3 \quad (19)$$

$$\min_{t_i} \sum_{e \in E(i,j)} \rho_t(\|R_e^{in} t_i + t_e^{in} - t_j\|) \quad \text{s.t. } t_1 = 0 \quad (20)$$

where ρ_R and ρ_t are Geman-Mcclure robust functions:

$$\rho_R(x) = \frac{s_R^2 x^2}{s_R^2 + x^2}, \quad \rho_t(x) = \frac{s_t^2 x^2}{s_t^2 + x^2} \quad (21)$$

where the hyper-parameters s_R and s_t are given by

$$s_R = 2 \cdot \underset{1 \leq i \leq N}{\text{median}} \min_{j : E(i,j) = \emptyset} \|R_e^{in} - R_j^g R_i^{gT}\|_F$$

$$s_t = 2 \cdot \underset{1 \leq i \leq N}{\text{median}} \min_{j : E(i,j) = \emptyset} \|R_e^{in} t_i^g + t_{ij}^{in} - t_j^g\|$$

Here (R_i^g, t_i^g) are the optimal scan poses from step II.

Unlike [4] that solves (19) and (20) using iteratively reweighted non-linear least squares (IRNLLS), we solve both using Gauss-Newton method (GNM). This is motivated from the fact that the output of step II is already close to the optimal solution. Specifically, starting from $R_i^{(1)} = R_i^g$, at iteration k we optimize the velocity vector $c_i \in R^3$ for R_i by minimizing

$$F = \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \sum_{e \in E(i,j)} \frac{s_R^2}{s_R^2 + \|E_e^{(k)}\|^2} \|E_e^{(k)} - (c_j \times) R_{ij}^{(k)} x + R_{ij}^{(k)} (c_i \times) - \frac{E_e^{(k)} \langle E_e^{(k)}, -(c_j \times) R_{ij}^{(k)} + R_{ij}^{(k)} (c_i \times) \rangle}{s_R^2 + \|E_e^{(k)}\|_F^2}\|_F^2,$$

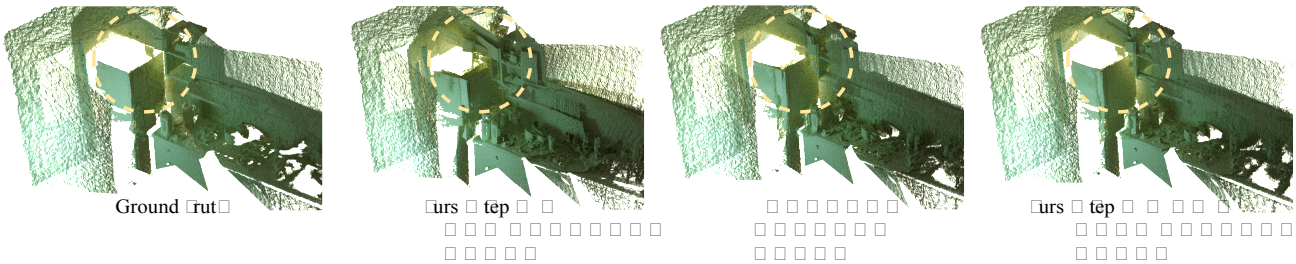

Figure 4. Comparison between reconstruction results before and after GNM for optimizing (19) on a dataset with 18 scans, with mean rotation and translation errors below the visualization.

s.t $c_1 = 0$, where $R_{ij}^{(k)} = R_j^{(k)} R_i^{(k)T}$ and $E_e^{(k)} = R_e^{in} - R_{ij}^{(k)}$. F is quadratic in c_i, therefore, the optimal values c_i^* can be obtained by solving a linear system. After that, we update the rotations as $R_i^{(k+1)} = \exp(\alpha^* c_i^* \times) R_i^{(k)}$, where α^* is the optimal stepsize obtained through line search.

Similarly, starting from $t_i^{(1)} = t_i^g$, at iteration k we optimize the displacement vector $\bar{c}_i \in \mathbb{R}^3$ for t_i via

$$\min_{\bar{c}_i} \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \sum_{e \in E(i,j)} \frac{\sigma_t^2}{\sigma_t^2 + \|r_e^{(k)}\|^2} \left\| r_e^{(k)} + R_{ij}^{in} \bar{c}_i - \bar{c}_j \right\|^2 - \frac{r_e^{(k)} r_e^{(k)T}}{\sigma_t^2 + \|r_e^{(k)}\|^2} R_e^{in} \bar{c}_i - c_j \Big\|^2 \quad \text{s.t. } \bar{c}_1 = 0 \tag{22}$$

where $r_e^{(k)} = R_e^{in} t_i^{(k)} + t_e^{in} - t_j^{(k)}$. Figure 4 shows the significant improvement in synchronization accuracy with GNM on a dataset with 18 scans. Please refer to the supp. material for more comparisons.

4. Evaluation

Section 4.1 introduces the experimental setup. Section 4.2, Section 4.3, and Section 4.4 present the results on synthetic datasets, RGB-D scans, and RGB images, respectively.

4.1. Experimental Setup

Baseline approaches. We consider four types of baseline approaches.

- Convex programming relaxation. We evaluate against SE-Sync [28], a state-of-the-art approach for pose synchronization. We also compare the performance of rotation synchronization against two other approaches: Shonan-Rot [11] and SDP-Rot [34].
- Spectral synchronization. The second category of baseline employs spectral approaches to perform synchronization. We compare against TransSync [20], a state-of-the-art approach in this category.
- Non-convex optimization. The third category of baselines minimizes non-convex objective functions. We compare against SFM-MRF [9] and K-Best [33], two state-of-the-art approaches, and IRLS-L0 [5], a state-of-the-art rotation synchronization approach.
- Cycle-pruning. The fourth category leverages the cycle-consistency constraint to filter out incorrect relative poses. We compare against RobustRecons [6], a state-of-the-art approach in this category.

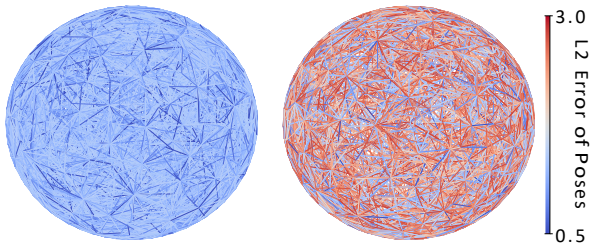

Figure 5. Synthetic datasets where G are geometric graphs on 1000 random samples on the unit sphere. (Left) Sync-Easy. The average vertex degree is 30. (Right) Sync-Hard. The average degree is 20.

Evaluation metrics. Following [9], we report the percentage of absolute angular rotation errors and absolute translation errors that fall within a varying threshold.

4.2. Results on Synthetic Datasets

As illustrated in Figure 5, we generate synthetic datasets by constructing G as geometric graphs. This is motivated from the trial application of pose synchronization in 3D reconstruction from RGB-D or RGB images where each object is matched with adjacent objects. Specifically, we consider a Sync-Easy and a Sync-Hard dataset, where both of them have 1000 objects distributed on a unit sphere. Each dataset is parameterized by five parameters k_n, n_g, p, q, and δ. k_n denotes the average number of adjacent objects in G. For Sync-Easy, we connect each object with 30 nearest neighbors. For Sync-Hard, we connect each object with 10 nearest neighbors. n_g denotes the number of absolute poses associated with each object, i.e., (R_i^l, t_i^l), $1 \le i \le n, 1 \le l \le n_g$. Here the ground-truth is given by $R_i^1 = I_3$, $t_i^1 = 0$. $R_i^l, 2 \le l \le n_g$ is a random rotation and $t_i^l, 2 \le l \le n_g$ is a random translation in the cube $[-1, 1]^3$. p, q, and δ control how the relative poses are generated. We generate

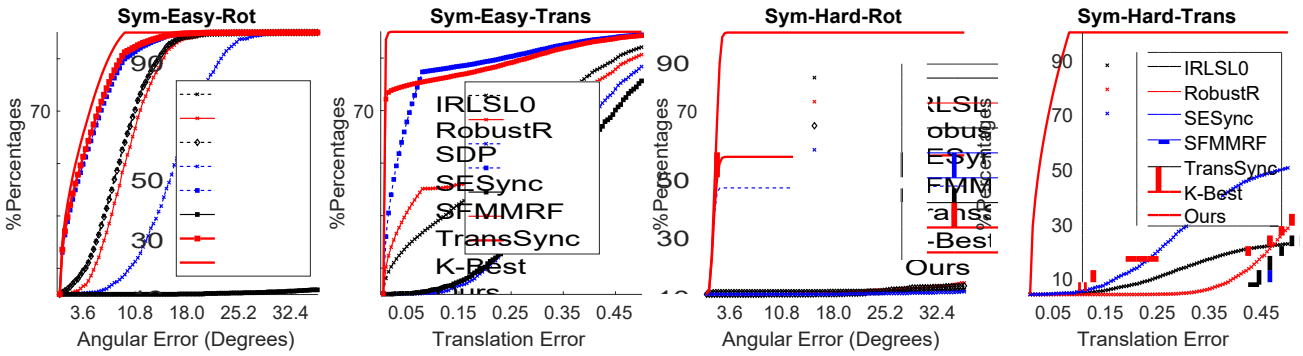

Figure 6. Results on synthetic datasets shown in Figure 5. (a) Angular rotation errors on the Sync-Easy dataset. (b) Translation errors on the Sync-Easy dataset. (c) Angular rotation errors on the Sync-Hard dataset. (d) Translation errors on the Sync-Hard dataset.Baselines: IRLSL0 [5]; RobustR [6]; SDP [34]; SESync [28]; SFMMRF [9]; TransSync [20];K-Best [33].

n_g relative poses along each edge: $(R^l_{ij}, t^l_{ij}), 1 \le l \le n_g$. With probability p_l,

$$R^l_{ij} = \exp(c_R^{ijl}\times) R^l_j {R^l_i}^T, t^l_{ij} = t^l_j - R^l_j {R^l_i}^T t^l_i + c_t^{ijl}$$

where c_R^{ijl} and c_t^{ijl} are independent random vectors in cube $[-\delta, \delta]^3$. Likewise, with probability $1 - p_l$, R^l_{ij} is a random rotation and t^l_{ij} is a random vector in cube $[-1, 1]^3$. p_l are chosen so that $p_1 = p$ and $p_l = q, 2 \le l \le n_g$. The parameters for Sync-Easy and Sync-Hard are $k_n = 30, n_g = 2, p = 1, q = 0.5, \delta = 4\cdot 10^{-3}$ and $k_n = 20, n_g = 3, p = 0.8, q = 0.5, \delta = 2\cdot 10^{-2}$.

Figure 6 compares our approach and baseline approaches. First of all, we can see that on both Sync-Easy and Sync-Hard, convex optimization formulations (SE-Sync, Shonan-Rot, and SDP-Rot) and spectral formulations (TransSync) fail to recover the underlying ground-truth. This is expected as both formulations are based on the assumption that there is at most one relative pose estimation between each pair of objects. RobustRecons and IRLS-L0 deliver reasonable results on Sync-Easy. We can understand this from the fact that pruning inconsistent cycles applies when the noise-level among inliers is small, while optimization under robust norm may work when the fraction of inliers is significant (50% for Sync-Easy). However, both RobustRecons and IRLS-L0 fail on Sync-Hard, in which cycle-consistency is not exactly satisfied among inliers and the fraction of outliers is significant (80% for Sync-Hard).

The top performing baselines are SFM-MRF and K-Best, which can take as input multiple relative poses between a pair of objects. However, the accuracy of our approach is much higher than both. This is because both baselines sample the space of rotations and translations, which is subject to discretization errors. Such discretization errors affect the optimal solution greatly when the fraction of outliers is large. Therefore, our approach outperforms SFM-MRF and K-Best significantly on Sym-Hard.

4.3. Results on RGB-D Scans

Next, we compare our approach and baseline approaches on pose synchronization of RGB-D scans. To this end, we use ScanNet [10], which consists of 1513 3D scenes. Our goal is to test the performance of 3D reconstruction from sparse views. To generate the input data, we randomly sample 10 to 25 scans from each scene. The relative poses between each pair of scans are computed using spectral matching [35]. We consider two settings. In the first setting, we consider one relative pose between each object pair whenever the number of matched features is bigger than 20% times the minimum number of features on each object. In the second setting, we use up to four relative poses per object pair whenever the number of matched features is bigger than 20% times the minimum number of features on each object. We report statistics over the entire ScanNet dataset.

Figure 7(Left) shows the rotation and translation errors of our approach and baseline approaches under the first setting. In this setting, the major challenge is that most edges are outliers that do not have a reasonable relative pose estimate. In this case, all methods, including our approach, do not produce accurate synchronization results. However, our approach outperforms all baseline approaches when considering the relative performance with salient margins. Specifically, regarding the percentage of rotation estimates whose errors fall under 10°, our approach achieves 56.4%, which is 17.3% higher than K-Best, the top-performing baseline. Regarding the percentage of translation estimates whose errors fall under 0.25m, our approach achieves 58.9%, which is again 4.6% higher than K-Best.

Figure 7(Right) shows the rotation and translation errors of our approach and baseline approaches under the second setting. Due to representation limitations of convex programming and spectral approaches, i.e., SE-Sync, Shanon-Rot, SDP-Rot, and TransSync, their performance is worse than the first setting. Thanks to certain abilities for handling multiple relative poses per object pair, IRLS-L0, RobustRecons, SFM-MRF, and K-

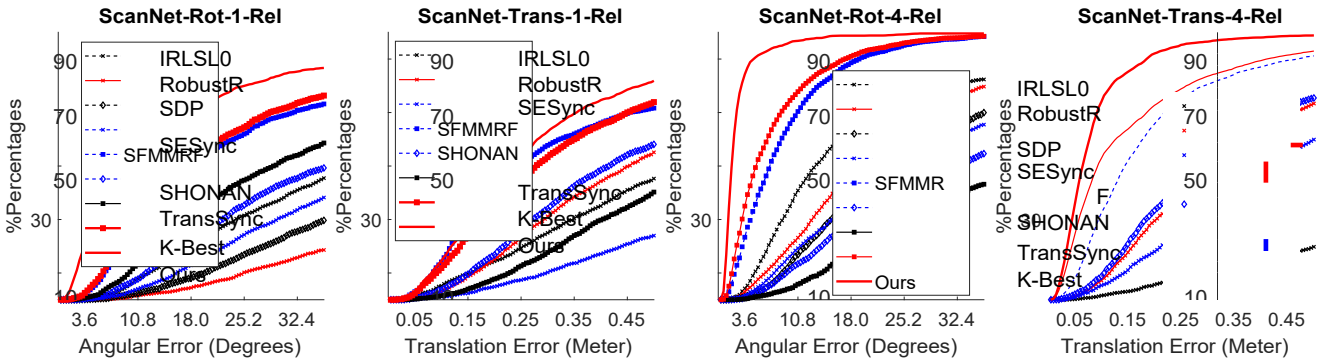

Figure 7. Results on ScanNet [10]. (Left) Average angular rotation errors (a) and average translation errors given one relative pose per object pair. (Right) Average angular rotation errors (c) and average translation errors (d) given four relative poses per object pair.

Best result in performance improvements compared to the first setting. In particular, K-Best is still the top-performing baseline with 67.8% percent of rotation predictions under 10° and 73.2% percent of translation predictions under 0.25m. Our approach is significantly better than all baseline approaches, improving the rotation and translation estimates by 22.1% and 12.5% under the same thresholds. Moreover, the performance gains of our approach in the second setting are larger than in the first setting. In other words, our approach can fully utilize the additional inputs.

4.4. Results on RGB Images

Finally, we evaluate rotation synchronization on RGB images. To this end, we consider two large-scale datasets Cornell-Artquad and San-Francisco from [9]. Cornell-Artquad has 5530 images with 222044 edges. San-Francisco has 7866 images and 101512 edges. Each edge has one relative rotation as input.

Figure 8 plots the angular errors of our approach and baseline approaches. Since G is sparse, approaches based on low-rank matrix recovery that require noise-level smaller than the spectral gap of G, i.e., SE-Sync, Shonan-Rot, SDP-Rot, and Trans-Sync, only exhibit limited performance. For the same reason, the performance of these approaches on San-Francisco is lower than that on Cornell-Artquad. RobustRecons also shows limited performance on both datasets due to difficulty in sampling cycles on sparse graphs. The top-performing baselines are SFM-MRF and K-Best. However, our approach still outperforms both of them. The running time of our method is 263s over San-Francisco and 570s over Cornell-Artsquad, both of which are competitive against baseline methods (see Appendix C for a full comparison of time efficiency).

5. Limitations

One limitation of our approach is that candidate pose generation employs pose clustering at each iteration for each object. When the number of adjacent edges is large, this step is time-consuming. We plan to address this challenge by developing more efficient clustering

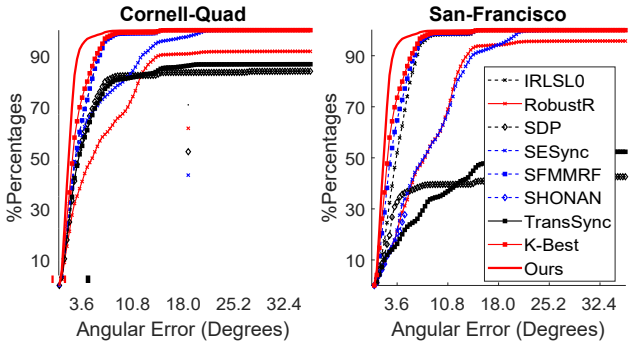

Figure 8. Results of rotation synchronization on (Left) Cornell-Artquad and (Right) San-Francisco.

approaches or using a subset of propagated poses for clustering. Another limitation is that step I which computes candidate poses, and step II, which jointly selects candidate poses, are decoupled. We plan to study how to unify these two steps as both steps employ a diffusion operation that aggregates information from neighboring objects.

6. Conclusion

This paper introduced a new pose synchronization approach that addresses the open challenge where there are multiple relative poses between each pair of objects, in which the correct one may not be the top ranked relative pose. Our approach employs a novel iterative diffusion and clustering step for computing the candidate poses, a joint optimization strategy for selecting consistent poses, and a pose refinement step which optimizes the relative poses under a robust norm. Empirical results on real and synthetic datasets demonstrate the advantage of our approach against state-of-the-art pose synchronization approaches.

Acknowledgement. We would like to acknowledge NSF IIS-2047677, HDR-1934932, and CCF-2019844.

References

[1] Federica Arrigoni, Andrea Fusiello, and Beatrice Rossi. Camera motion from group synchronization. In 3D Vision (3DV), 2016 Fourth International Conference on, pages 546–555. IEEE, 2016.

[2] Federica Arrigoni, Beatrice Rossi, and Andrea Fusiello. Spectral synchronization of multiple views in se (3). SIAM Journal on Imaging Sciences, 9(4):1963–1990, 2016.

[3] Florian Bernard, Johan Thunberg, Peter Gemmar, Frank Hertel, Andreas Husch, and Jorge Goncalves. A solution for multi-alignment by transformation synchronisation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2161–2169, 2015.

[4] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In ICCV, pages 521–528. IEEE Computer Society, 2013.

[5] Avishek Chatterjee and Venu Madhav Govindu. Robust relative rotation averaging. IEEE transactions on pattern analysis and machine intelligence, 40(4):958–972, 2017.

[6] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In CVPR, pages 5556–5565. IEEE Computer Society, 2015.

[7] Trial Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell., 24(5):603–619, may 2002.

[8] David Crandall, Andrew Owens, Noah Snavely, and Daniel Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 35(12):2841–2853, December 2013.

[9] David J. Crandall, Andrew Owens, Noah Snavely, and Daniel P. Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. IEEE Trans. Pattern Anal. Mach. Intell., 35(12):2841–2853, 2013.

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. cite arxiv:1702.04405.

[11] Frank Dellaert, David M Rosen, Jing Wu, Robert Mahony, and Luca Carlone. Shonan rotation averaging: Global optimality by surfing so(n). In European Conference on Computer Vision, pages 292–308. Springer, 2020.

[12] Leonidas J. Guibas, Qixing Huang, and Zhenxiao Liang. A condition number for joint optimization of cycle-consistent networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 1005–1015, 2019.

[13] Qi-Xing Huang, Yuxin Chen, and Leonidas J. Guibas. Scalable semidefinite relaxation for maximum A posterior estimation. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 64–72, 2014.

[14] Qixing Huang, Simon Flöry, Natasha Gelfand, Michael Hofer, and Helmut Pottmann. Reassembling fractured objects by geometric matching. In ACM SIGGRAPH 2006 Papers, SIGGRAPH '06, pages 569–578, New York, NY, USA, 2006. ACM.

[15] Qixing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing, pages 177–186, 2013.

[16] Qixing Huang, Fan Wang, and Leonidas J. Guibas. Functional map networks for analyzing and exploring large shape collections. ACM Trans. Graph., 33(4):36:1–36:11, 2014.

[17] Qi-Xing Huang, Guo-Xin Zhang, Lin Gao, Shi-Min Hu, Adrian Butscher, and Leonidas J. Guibas. An optimization approach for extracting and encoding consistent maps in a shape collection. ACM Trans. Graph., 31(6):167:1–167:11, 2012.

[18] Qi-Xing Huang, Guo-Xin Zhang, Lin Gao, Shi-Min Hu, Adrian Butscher, and Leonidas Guibas. An optimization approach for extracting and encoding consistent maps in a shape collection. ACM Trans. Graph., 31(6), nov 2012.

[19] Xiangru Huang, Zhenxiao Liang, Chandrajit Bajaj, and Qixing Huang. Translation synchronization via truncated least squares. In NIPS, 2017.

[20] Xiangru Huang, Zhenxiao Liang, Xiaowei Zhou, Yao Xie, Leonidas J Guibas, and Qixing Huang. Learning transformation synchronization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8082–8091, 2019.

[21] Daniel Huber. Automatic Three-dimensional Modeling from Reality. PhD thesis, Robotics Institute , Carnegie Mellon University, Pittsburgh, PA, December 2002.

[22] Jörg H. Kappes, Björn Andres, Fred A. Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X. Kausler, Jan Lellmann, Nikos Komodakis, and Carsten Rother. A comparative study of modern inference techniques for discrete energy minimization problems. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, pages 1328–1335. IEEE Computer Society, 2013.

[23] Vladimir G. Kim, Wilmot Li, Niloy J. Mitra, Stephen DiVerdi, and Thomas Funkhouser. Exploring collections of 3d models using fuzzy correspondences. ACM Trans. Graph., 31(4):54:1–54:11, July 2012.

[24] Spyridon Leonardos, Xiaowei Zhou, and Kostas Daniilidis. Distributed consistent data association via permutation synchronization. In ICRA, pages 2645–2652. IEEE, 2017.

[25] Shaohan Li, Yunpeng Shi, and Gilad Lerman. Fast, accurate and memory-efficient partial permutation synchronization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 15714–15722. IEEE, 2022.

[26] Deepti Pachauri, Risi Kondor, Gautam Sargur, and Vikas Singh. Permutation diffusion maps (PDM) with application to the image association problem in computer vision. In NIPS, pages 541–549, 2014.

[27] Deepti Pachauri, Risi Kondor, and Vikas Singh. Solving the multi-way matching problem by permutation synchronization. In NIPS, pages 1860–1868, 2013.

[28] D.M. Rosen, L. Carlone, A.S. Bandeira, and J.J. Leonard. SE-Sync: A certifiably correct algorithm for synchronization over the special Euclidean group. Intl. J. of Robotics Research, 38(2–3):95–125, Mar. 2019.

[29] Yanyao Shen, Qixing Huang, Nati Srebro, and Sujay Sanghavi. Normalized spectral map synchronization. In NIPS, pages 4925–4933, 2016.

[30] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. Applied and computational harmonic analysis, 30:20–36, 2011.

[31] Amit Singer and Hau tieng Wu. Vector diffusion maps and the connection laplacian. Communications on Pure and Applied Mathematics, 65:1067–1144, March 2012.

[32] Yifan Sun, Zhenxiao Liang, Xiangru Huang, and Qixing Huang. Joint map and symmetry synchronization. In Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V, pages 257–275, 2018.

[33] Yifan Sun, Jiacheng Zhuo, Arnav Mohan, and Qixing Huang. K-best transformation synchronization. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 10251–10260, 2019.

[34] Lanhui Wang and Amit Singer. Exact and stable recovery of rotations for robust synchronization. Information and Inference: A Journal of the IMA, 2:145–193, December 2013.

[35] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2452–2461, 2020.

[36] Yuxin Chen and Emmanuel Candes. The projected power method: An efficient algorithm for joint alignment from pairwise differences. Communications on Pure and Applied Mathematics, 71(8), August 2018.

[37] Yuxin Chen, Leonidas J. Guibas, and Qi-Xing Huang. Near-optimal joint object matching via convex relaxation. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 100–108, 2014.

[38] Christopher Zach, Manfred Klopschitz, and Manfred Pollefeys. Disambiguating visual relations using loop constraints. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1426–1433, 2010.

[39] Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In Proceedings of the IEEE International Conference on Computer Vision, pages 4032–4040, 2015.