SIM-Sync-Mono: Joint Depth Estimation and Certifiably Optimal Synchronization Using Learned Module

Xihang Yu
University of Michigan
xihangyu@umich.edu

Guoyuan Li
University of Michigan
secondauthor@i2.org

Yuchen Zhou
University of Michigan
secondauthor@i2.org

Abstract

Building on [40], this study introduces an innovative approach for estimating camera trajectories and 3D scene structures from multiview image keypoints, utilizing a pretrained depth prediction network. Our method sets itself apart from previous methods [22, 20, 41] by efficiently separating camera pose estimation and depth fine-tuning. We employ a SIM-Sync *solver to optimally solve the structured problem of camera trajectory estimation, while depth fine-tuning, a more complex task, is addressed using nonlinear solvers. This strategic separation not only exploits the structures in camera trajectory estimation but also simplifies the optimization process, especially in iterative refinement within video sequences. Additionally, our approach benefits from a more straightforward and efficient loss function design, enhancing the overall effectiveness of the method. The key contributions of this work include (i) the development of a unified solver for both camera trajectory and depth fine-tuning, (ii) validated through experiments on the* TUM *dataset, and (iii) an interactive demonstration available on Google Colab.*

1. Introduction

Building on [40], a previous work that offers a *certifiably optimal* solution for estimating camera trajectory and 3D scene structure *directly from multiview image keypoints*, this project addresses the gap between pose graph optimization and bundle adjustment. While the former allows efficient global optimization with relative pose measurements, the latter, though it directly utilizes image keypoints, faces challenges in global optimization due to the complexity of camera projective geometry. The solution presented bridges this gap through a *pretrained* depth prediction network. In this approach, nodes in a graph represent monocular images captured at unknown camera poses, and edges indicate pairwise image keypoint correspondences. SIM-Sync employs a pretrained depth network to *lift* 2D keypoints into 3D *scaled* point clouds, contending with scale ambiguity inherent in monocular depth prediction. The goal of SIM-Sync is to *synchronize* the unknown camera poses and scaling factors (*i.e.* over the 3D similarity group) by minimizing the Euclidean distances between scaled point clouds. This formulation of SIM-Sync, although nonconvex, facilitates the design of a certifiably optimal solver akin to the SE-Sync algorithm. It tackles translations in a closed-form manner, while the optimization of rotations and scales transforms into a *quadratically constrained quadratic program*. Here, Shor's semidefinite relaxation technique is applied, with scale regularization integrated into the semidefinite program to avoid scale estimation contraction. A graphical representation of SIM-Sync, exemplified using the TUM dataset [31], is depicted in Fig. 1.

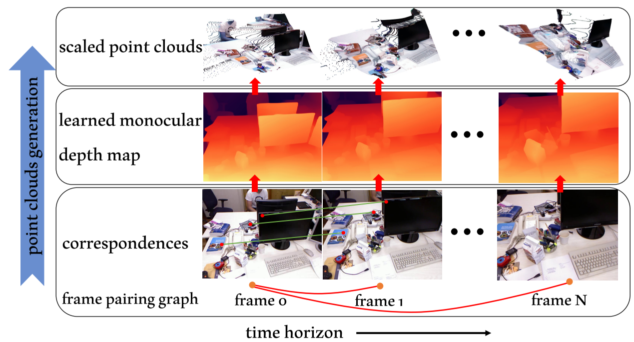

Figure 1. Illustration of SIM-Sync on the TUM dataset [31]. Feature matching algorithms extract correspondences, which are then elevated using pre-trained depth networks to generate scaled point clouds. These point clouds are integrated with the frame graph to optimize the camera trajectory effectively.

When the authors conducted research in SIM-Sync, they discovered that the performance largely depend on accuracy of depth prior. So, our question is *how to leverage the 3D reconstruction from* SIM-Sync *to improve the imperfect depth prediction from the pretrained model*. Our project introduces a novel methodology for dense depth estimation and 3D camera trajectory in video sequences, optimizing camera trajectory and depth networks jointly. Camera depth finetune can benefit from two perspectives: From optimization perspective, we can repeatedly alternate the camera trajectory estimation and depth finetuning. Ideally, both camera trajectory and depth estimation converge to the optimal value. From transfer learning's perspective, pretrained depth model (a model trained on a large dataset) may not be accurate for specific environment. However, finetuned model is expected to perform better for unseen sequence in the same environment. Similar ideas have been proposed in [22, 20, 41]. However, this approach diverges from recent trends that fine-tune pretrained networks on input videos [41, 20, 22], especially in its departure from using back-propagation on complex, nonlinear loss functions. [20, 22] optimizes camera trajectory and but does not finetune the depth network. [41] can alternatively optimizes camera trajectory and finetunes depth network. However, it hand designs very complex loss functions to regulate both camera trajectory estimation and depth finetune step. Instead, we delineate camera pose estimation and depth fine-tuning. Camera pose estimation is addressed with the SDP solver SIM-Sync, ensuring global optimality. The motivation of the process is divide-and-conquer strategy. This strategy exploits structural advantages in camera trajectory estimation by solving it optimally, while relegating the more challenging depth fine-tuning to nonlinear solvers. Our approach also benefits from simpler loss function designs compared to previous methods.

Contributions. This project offers three contributions:

- Introduction of SIM-Sync-Mono, a unified solver that simultaneously addresses camera trajectory estimation and depth fine-tuning.
- Experimental validation using the TUM dataset.
- An interactive open-sourced demonstration on Google Colab.

2. Related Works

We review related work on structure from motion and visual SLAM in Section 2.1, and on certifiable geometric perception in Section 2.2.

2.1. Structure from Motion and SLAM

Estimating camera poses and scene structure from sensor data is a long-standing problem in computer vision and robotics. This problem is variously called structure from motion (SfM) [28] or simultaneous localization and mapping (SLAM) [7], where SLAM can often rely on GPS, IMU, and even wireless communication [18]. In classic SfM and SLAM, the problem is typically decomposed into feature matching and geometric estimation, where feature matching establishes keypoint correspondences between images (and point clouds) and geometric estimation seeks to find the best poses and structure that fit the correspondences. Feature matching, closely related to representation learning, is one of the most popular topics in computer vision with a vast amount of literature, for which we refer to [1] for a recent review. When it comes to geometric estimation, as introduced in Section 1, two popular paradigms are the pose graph optimization formulation in SLAM and the bundle adjustment formulation in SfM.

Recently, a number of methods seek to integrate learned components into classic SfM or SLAM methods. [23] explored jointly optimizing depth maps, camera poses and confidence masks for weighting the photo-metric loss during training. [43] learns depth and ego-motion from monocular video without supervision. [39] leverages modules for learned prediction of depth, pose, and uncertainty within a bundle adjustment framework. DROID-SLAM [32] makes use of a dense bundle adjustment layer to update depth map and camera poses concurrently. [25] leverages pre-trained depth prediction for better initialization of visual inertial odometry.

In this work, we use a pretrained depth prediction network to lift 2D image keypoints as 3D scaled point clouds to enable global synchronization of camera poses and unknown scaling coefficients in depth prediction.

2.2. Certifiably Optimal Geometric Perception

Certifiably optimal geometric perception refers to developing algorithms that either solve geometric estimation problems to global optimality and produce an optimality certificate, or fail to do so but provide a bound of sub-optimality [37, Definition 1]. Semidefinite programming has been the major tool for developing certifiably optimal estimation algorithms. The pioneering work by Kahl and Henrion [19] employs Lasserre's hierarchy to tackle various early perception problems, including camera resectioning, homography estimation, and fundamental matrix estimation. More recently, certifiable algorithms have been developed for modern applications such as outlier-robust estimation [37, 36], pose graph optimization [9, 27], rotation averaging [12, 13], triangulation [3, 11], 3D registration [38, 5, 10, 17, 24], absolute pose estimation [2], relative pose estimation [6, 14, 42], hand-eye calibration [15, 16, 34], uncertainty propagation in non-rigid SfM [30] and category-level object perception [29, 35].

In this work, we develop the first certifiably optimal algorithm that estimates 3D scene and camera poses directly from 2D image correspondences.

3. Methods

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V} = [N]$ is associated with an RGB image $I_i \in \mathbb{R}^{H \times W \times 3}$ and an unknown camera pose $(R_i, t_i) \in \mathrm{SE}(3)$, and each edge

$(i,j) \in \mathcal{E}$ contains a set of n_{ij} dense pixel-to-pixel correspondences $\mathcal{C}_{ij} = \{p_{i,k} \leftrightarrow p_{j,k}\}_{k=1}^{n_{ij}}$ with $p_{i,k} \in \mathbb{R}^2$ the k-th pixel location in image I_i and $p_{j,k} \in \mathbb{R}^2$ the k-th pixel location in image I_j. Assuming all the camera intrinsics $\{K_i\}_{i=1}^N$ are known, we can compute

$$\widetilde{p}_{i,k} = K_i^{-1} \begin{bmatrix} p_{i,k}^x & p_{i,k}^y & 1 \end{bmatrix}^\mathsf{T} \tag{1}$$

as the *bearing vector* normalized by the camera intrinsics. The third entry of $\widetilde{p}_{i,k}$ is equal to 1.

Pretrained depth prediction. Suppose we are given a pretrained depth estimation network that, for each image I_i, produces a depth map. Let $d_{i,k} > 0$ be the predicted depth of $p_{i,k}$ and $s_i > 0$ be the unknown scale coefficient for image I_i. Consequently, $\widehat{p}_{i,k} = s_i d_{i,k} \widetilde{p}_{i,k}$ corresponds to the 3D location of $p_{i,k}$ in the i-th camera frame. Effectively, with the depth predictor, for every $(i,j) \in \mathcal{E}$, we have a pair of *scaled* point cloud measurements $\{d_{i,k}\widetilde{p}_{i,k}\}_{k=1}^{n_{ij}}$ and $\{d_{j,k}\widetilde{p}_{j,k}\}_{k=1}^{n_{ij}}$, as shown in Fig. 1.

The SIM-Sync formulation. We are interested in estimating the unknown camera poses and the per-image scale coefficients $\{x_i = (s_i, R_i, t_i)\}_{i=1}^N$. We formulate the following optimization

$$\min_{\substack{s_i>0, R_i\in \mathrm{SO}(3), t_i \in \mathbb{R}^3 \\ i=1,\dots,N}} \sum_{(i,j)\in\mathcal{E}} \sum_{k=1}^{n_{ij}} w_{ij,k} \left\| r_{ij,k} \right\|^2 \tag{SIM-Sync}$$

with

$$r_{ij,k} = (R_i(s_i d_{i,k}\widetilde{p}_{i,k}) + t_i) - (R_j(s_j d_{j,k}\widetilde{p}_{j,k}) + t_j)\,.$$

The objective tries to minimize the 3D point-to-point distances as (R_i, t_i) transforms $\widehat{p}_{i,k}$, and (R_j, t_j) transforms $\widehat{p}_{j,k}$ into the same global coordinate frame. In (SIM-Sync), we include $w_{ij,k} > 0$ for generality: these known weights capture the potential uncertainty of the correspondences. Usually these weights are unknown and in our experiments we use GNC and TEASER to estimate them so that $w_{ij,k} = 1$ indicates inliers and $w_{ij,k} = 0$ indicates outliers.

Anchoring. Problem (SIM-Sync) is ill-defined. One can choose $s_i \to 0, \forall i = 1,\dots,N$, $t_1 = t_2 = \dots = t_N =$ constant, and the objective of (SIM-Sync) can be set arbitrarily close to zero. To resolve this issue, we anchor the first frame and set $R_1 = \mathbf{I}_3, t_1 = \mathbf{0}, s_1 = 1$, which is common practice in many related pose graph estimation formulations [27].

Camera Depth Finetune: In the second stage, we formulate (Depth). Note that the objective function is the same as in (SIM-Sync). However, decision variable changes to depth parameters. Compared with [22, 20, 41], we hope SDP reformulation for camera trajectory estimation can improve the efficiency and accuracy of camera trajectory estimation and depth finetune.

$$\min_{\substack{d_{i,k}>0, d_{j,k}>0 \\ (i,j)\in\mathcal{E}, k=1,\dots,n_{ij}}} \sum_{(i,j)\in\mathcal{E}} \sum_{k=1}^{n_{ij}} w_{ij,k} \left\| r_{ij,k} \right\|^2 \tag{Depth}$$

with

$$r_{ij,k} = (R_i(s_i d_{i,k}\widetilde{p}_{i,k}) + t_i) - (R_j(s_j d_{j,k}\widetilde{p}_{j,k}) + t_j)\,.$$

Algorithm 1: SIM-Sync-Mono Algorithm

```
Input: I, Π_0
Initialize: N Number of iterations
for i = 1, ..., N do
    E ← EssentialGraphRetriever(I)
    D ← Π_i(I, E)
    P ← PointCloudRetriever(D, I, E)
    R_i, t_i, s_i ← SIM-Sync(E, P)
    Π_i ← DepthFinetuner(R_i, t_i, s_i)
end
return (R_N, t_N, s_N, Π_N)
```

In Algorithm 1, we process image sequences I and utilize a pre-trained depth network Π_0. Initially, we set the iteration count N. Each iteration involves acquiring edges E via ORB-SLAM3 [8], and initial correspondences are established using SIFT [21]. The CAPS descriptor [33] then refines these correspondences, limited to 400 from SIFT, based on feature similarities. Depths are derived from the pre-trained network [4], leading to the extraction of point clouds and edges. Inputs fed into SIM-Sync yield rotation, translation, and scale factors. The core innovation lies in iteratively refining the depth network by incorporating these factors into (Depth) and applying backpropagation to the head layers of the MiDaS-v3 network. The head network of MiDaS-v3 that features seven layers is as follows: a convolutional layer for feature extraction, an upsampling layer for spatial enlargement, another convolutional layer for further feature processing, a ReLU activation for non-linearity, followed by a third convolutional layer, a conditional ReLU or identity layer based on output requirements, and ending with an identity layer.

4. Experiments

Setup. We test two sequences in the TUM dataset, the first 200 frames in the `freiburg1_xyz` sequence and the first 200 frames in the `freiburg2_xyz` sequence, respectively.[1] For TEASER+SIM-Sync, we use learned depth obtained from the MiDaS-v3 model [26, 4], with the largest

[1] We discard the first 60 frames in `freiburg2_xyz` since the camera shakes and results in blurred images.

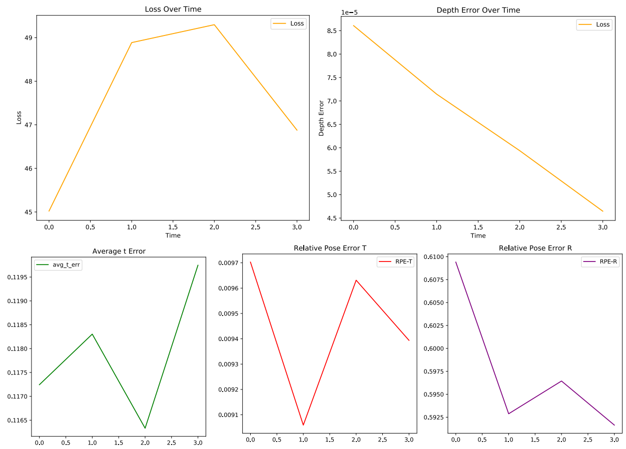

Figure 2. Illustration of SIM-Sync-Mono on the TUM dataset [31] `freiburg1_xyz` sequence.

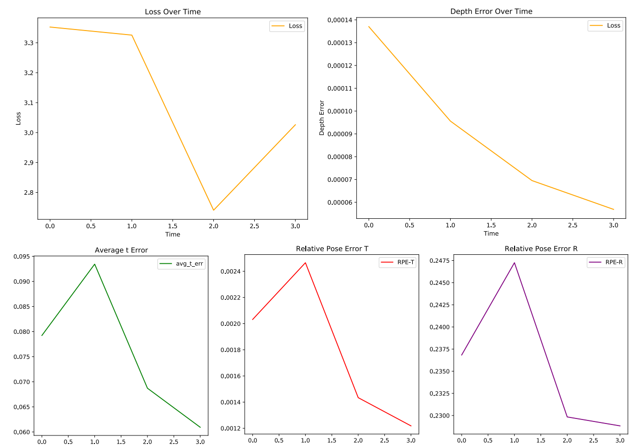

Figure 3. Illustration of SIM-Sync on the TUM dataset [31] `freiburg2_xyz` sequence.

10% depth discarded. Note that MiDaS-v3 is not trained on the TUM dataset, and we directly use its default parameter configuration (*i.e*. zero-shot). For number of iterations iterations, we set $N = 4$. We evaluate the finetuned depth estimation $\hat{d}_i$ against ground truth d_i using root mean square error:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left\|d_i - \hat{d}_i\right\|^2} \tag{2}$$

where M is number of all pixels in all images of a video. We follow the standard evaluation protocol of visual odometry for assessing pose accuracy, *i.e*. Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). ATE quantifies the root-mean-square error between predicted camera positions and the groundtruth positions. RPE measures the relative pose disparity between pairs of adjacent frames, including both translation error (RPE-T) and rotational error (RPE-R).

Results. The quantitative results, depicted in Fig. 2 and Fig. 3, illustrate that for the `freiburg1_xyz` sequence, SIM-Sync-Mono enhances depth and relative pose estimation, albeit with a slight decline in absolute translation accuracy, as indicated by the green line comparing the final results at iteration 4 to the initial at iteration 1. In contrast, the `freiburg2_xyz` sequence shows improvement across all metrics with SIM-Sync-Mono. Notably, both sequences exhibit fluctuations in pose estimation but demonstrate consistent advancements in depth estimation over time.

Interactive Colab. We have also open-sourced an interactive scripts for playing with SIM-Sync-Mono system. In this script, the user can run the blocks and get camera trajectory estimation and 3D reconstruction from bottom up with raw data.

5. Conclusion

In this study, we have presented a novel method that simultaneously fine-tunes depth estimation from a pretrained network and synchronizes camera trajectory, offering certifiable global optimality. Our experimental results affirm the method's effectiveness. Yet, unresolved issues persist. The choice of the optimal iteration count, N, is still open for determination. Furthermore, our validation is constrained to a pair of sequences from the TUM dataset, suggesting that adopting batched backpropagation could lead to further precision gains. Additionally, while depth fine-tuning has primarily been applied to optimize camera trajectory and depth predictions, its influence on transfer learning remains unexplored. Future efforts will involve partitioning the TUM dataset to distinguish between training and testing sets, thereby verifying the method's efficacy on new data and its ability to generalize without overfitting. A notable observation is that the finetuned depth, despite its increased accuracy, shows blurred edges, which may imply an overfit to the training sequences—an aspect that will be scrutinized in subsequent research.

References

[1] Image matching challenge 2023: The unbearable weight of the bundle adjustment. `https://ducha-aiki.github.io/wide-baseline-stereo-blog/2023/07/05/IMC2023-Recap.html`. Accessed: 2023-09-08.

[2] Sérgio Agostinho, João Gomes, and Alessio Del Bue. Cvxpnpl: A unified convex solution to the absolute pose estimation problem from point and line correspondences. *Journal of Mathematical Imaging and Vision*, 65(3):492–512, 2023.

[3] Chris Aholt, Sameer Agarwal, and Rekha Thomas. A qcqp approach to triangulation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence,*

Italy, October 7-13, 2012, Proceedings, Part I 12, pages 654–667. Springer, 2012.

[4] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

[5] Jesus Briales and Javier Gonzalez-Jimenez. Convex global 3d registration with lagrangian duality. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4960–4969, 2017.

[6] Jesus Briales, Laurent Kneip, and Javier Gonzalez-Jimenez. A certifiably globally optimal solution to the non-minimal relative pose problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–154, 2018.

[7] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016.

[8] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[9] Luca Carlone, Giuseppe C Calafiore, Carlo Tommolillo, and Frank Dellaert. Planar pose graph optimization: Duality, optimal solutions, and verification. *IEEE Transactions on Robotics*, 32(3):545–565, 2016.

[10] Kunal N Chaudhury, Yuehaw Khoo, and Amit Singer. Global registration of multiple point clouds using semidefinite programming. *SIAM Journal on Optimization*, 25(1):468–501, 2015.

[11] Diego Cifuentes. A convex relaxation to compute the nearest structured rank deficient matrix. *SIAM Journal on Matrix Analysis and Applications*, 42(2):708–729, 2021.

[12] Anders Eriksson, Carl Olsson, Fredrik Kahl, and Tat-Jun Chin. Rotation averaging and strong duality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 127–135, 2018.

[13] Johan Fredriksson and Carl Olsson. Simultaneous multiple rotation averaging using lagrangian duality. In *Asian Conference on Computer Vision*, pages 245–258. Springer, 2012.

[14] Mercedes Garcia-Salguero, Jesus Briales, and Javier Gonzalez-Jimenez. Certifiable relative pose estimation. *Image and Vision Computing*, 109:104142, 2021.

[15] Matthew Giamou, Ziye Ma, Valentin Peretroukhin, and Jonathan Kelly. Certifiably globally optimal extrinsic calibration from per-sensor egomotion. *IEEE Robotics and Automation Letters*, 4(2):367–374, 2019.

[16] Jan Heller, Didier Henrion, and Tomas Pajdla. Hand-eye and robot-world calibration by global polynomial optimization. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 3157–3164. IEEE, 2014.

[17] José Pedro Iglesias, Carl Olsson, and Fredrik Kahl. Global optimality for point set registration using semidefinite programming. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8287–8295, 2020.

[18] Ninad Jadhav, Weiying Wang, Diana Zhang, Oussama Khatib, Swarun Kumar, and Stephanie Gil. A wireless signal-based sensing framework for robotics. *The International Journal of Robotics Research*, 41(11-12):955–992, 2022.

[19] Fredrik Kahl and Didier Henrion. Globally optimal estimates for geometric reconstruction problems. *International Journal of Computer Vision*, 74:3–15, 2007.

[20] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[21] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[22] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020.

[23] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018.

[24] Haggai Maron, Nadav Dym, Itay Kezurer, Shahar Kovalsky, and Yaron Lipman. Point registration via efficient convex relaxation. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[25] Nathaniel Merrill, Patrick Geneva, and Saimouli Katragadda Chuchu Chen Guoquan Huang. Fast monocular visual-inertial initialization leveraging learned single-view depth. 2023.

[26] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[27] David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *The International Journal of Robotics Research*, 38(2-3):95–125, 2019.

[28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[29] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal pose and shape estimation for category-level 3d object perception. *arXiv preprint arXiv:2104.08383*, 2021.

[30] Jingwei Song, Mitesh Patel, Ashkan Jasour, and Maani Ghaffari. A closed-form uncertainty propagation in non-rigid structure from motion. *IEEE Robotics and Automation Letters*, 7(3):6479–6486, 2022.

[31] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

[32] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[33] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 757–774. Springer, 2020.

[34] Emmett Wise, Matthew Giamou, Soroush Khoubyarian, Abhinav Grover, and Jonathan Kelly. Certifiably optimal monocular hand-eye calibration. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 271–278. IEEE, 2020.

[35] Heng Yang and Luca Carlone. In perfect shape: Certifiably optimal 3d shape reconstruction from 2d landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 621–630, 2020.

[36] Heng Yang and Luca Carlone. One ring to rule them all: Certifiably robust geometric perception with outliers. *Advances in neural information processing systems*, 33:18846–18859, 2020.

[37] Heng Yang and Luca Carlone. Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2816–2834, 2022.

[38] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020.

[39] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020.

[40] Xihang Yu and Heng Yang. Sim-sync: From certifiably optimal synchronization over the 3d similarity group to scene reconstruction with learned depth. *arXiv preprint arXiv:2309.05184*, 2023.

[41] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.

[42] Ji Zhao. An efficient solution to non-minimal case essential matrix estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1777–1792, 2020.

[43] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.