Robust Reconstruction of Indoor Scenes

Sungjoon Choi*† Qian-Yi Zhou*‡ Vladlen Koltun‡

Abstract

We present an approach to indoor scene reconstruction from RGB-D video. The key idea is to combine geometric registration of scene fragments with robust global optimization based on line processes. Geometric registration is error-prone due to sensor noise, which leads to aliasing of geometric detail and inability to disambiguate different surfaces in the scene. The presented optimization approach disables erroneous geometric alignments even when they significantly outnumber correct ones. Experimental results demonstrate that the presented approach substantially increases the accuracy of reconstructed scene models.

1. Introduction

High-fidelity reconstruction of complete indoor scenes is known as a particularly challenging problem [19, 29, 64, 7]. Many indoor reconstruction systems make simplifying assumptions and forfeit detail in the reconstructed model [19, 64, 7], rely on user interaction [17], or both [37, 53]. Other systems rely on substantial hardware setups based on LiDAR scanners [11, 59].

The availability of consumer depth cameras provides an opportunity to develop robust reconstruction systems but does not in itself solve the associated challenges. While 3D models of real-world objects can now be created easily [46, 70], the same combination of quality and reliability has yet to be achieved for complete scenes. Unlike an object, which can be entirely in the field of view of the camera, a large scene must be reconstructed from views acquired along a complex trajectory, each view exposing only a small part of the environment. Camera paths that thoroughly image all surfaces at close range lead to significant odometry drift and the necessity to match and register different views globally.

Prior work on scene reconstruction with consumer depth cameras recognized the importance of global registration [29, 18, 69, 62, 65]. Nevertheless, no prior system appears to be sufficiently reliable to support automatic reconstruction of complete indoor scenes at a quality level appropriate for particularly demanding applications. This is evidenced by the recent effort of Xiao et al. to reconstruct a large number of indoor scenes. Due to the unreliability of automatic scene reconstruction pipelines, the authors resorted to manual labeling to establish correspondences between different views. ("existing automatic reconstruction methods are not reliable enough for our purposes." [65])

In this work, we present a fully automatic scene reconstruction pipeline that matches the reconstruction quality obtained with manual assistance by Xiao et al. and significantly exceeds the accuracy of prior automatic approaches to indoor reconstruction. An example reconstruction produced by our approach is shown in Figure 1. Our pipeline is geometric: pairs of local scene fragments are registered and a global model is constructed based on these pairwise alignments [31]. A critical weakness of such pipelines that we address is the low precision of geometric registration. Geometric registration algorithms are error-prone due to sensor noise, which leads to aliasing of fine geometric details and inability to disambiguate different locations based on local geometry. The difficulty is compounded by the necessity to register loop closure fragments that have low overlap. In practice, false pairwise alignments can outnumber correctly aligned pairs.

Our approach resolves inconsistencies and identifies correct alignments using global optimization based on line processes. Line processes were introduced in the context of image restoration as a means for automatically identifying discontinuities as part of a single global optimization [22, 21]. They are closely related to robust estimation [4]. The advantage of the line process formulation is that the optimization objective retains a least-squares form and can be optimized by a standard high-performance least-squares solver. We show that this framework is extremely effective in dealing with pairwise registration errors. Our implementation automatically prunes false pairwise alignments even when they significantly outnumber correct ones. Extensive experiments demonstrate that our approach substantially increases reconstruction accuracy.

Our work contains a number of supporting contributions of independent interest. First, we provide infrastructure for rigorous evaluation of scene reconstruction accuracy, augmenting the ICL-NUIM dataset [26] with challenging cam-

*Joint first authors
†Stanford University
‡Intel Labs

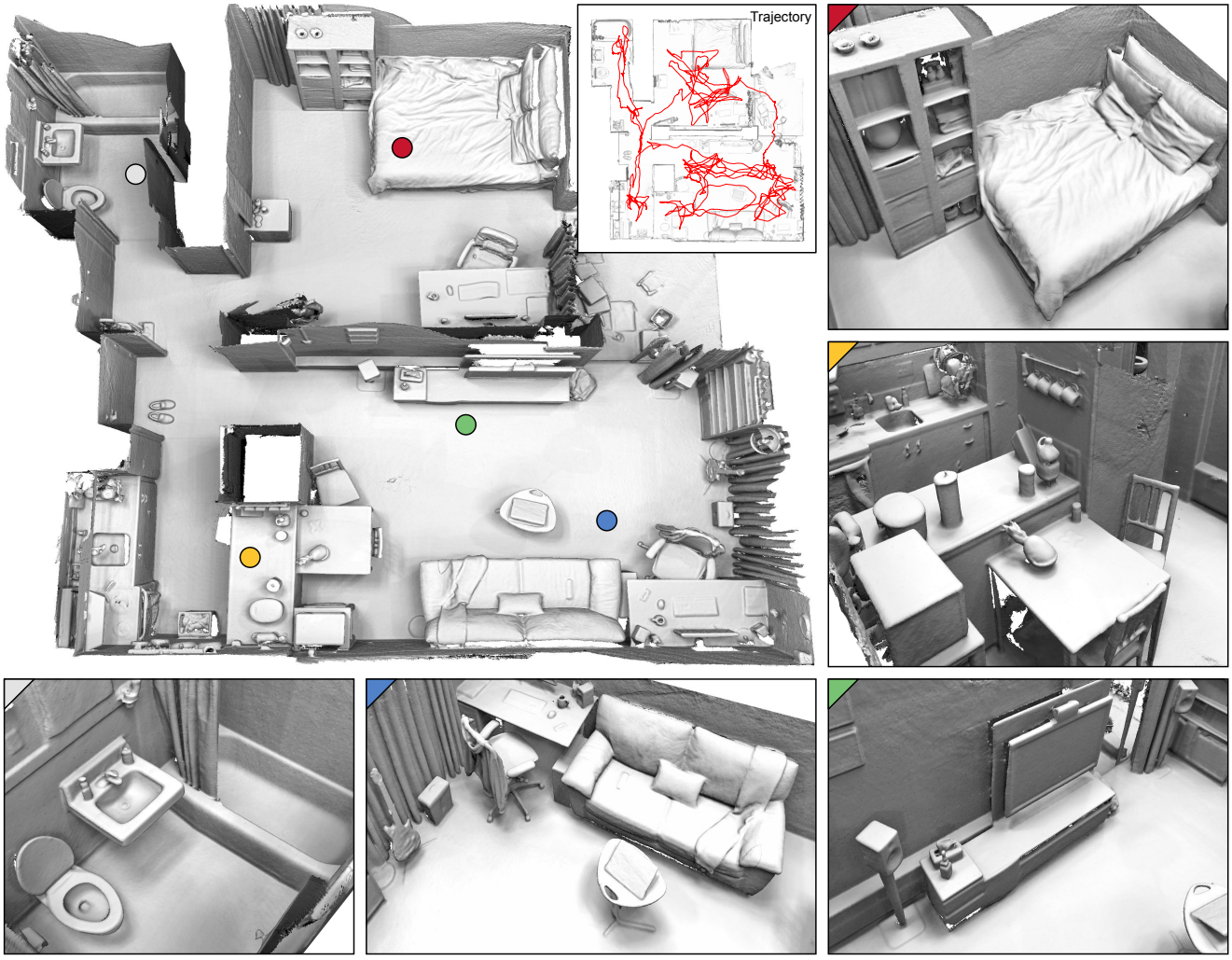


Figure 1. A complete apartment reconstructed by the presented approach. The estimated camera trajectory is 151.6 meters long, folded into a diameter of 8.3 meters.

era trajectories and a realistic noise model. Second, we perform a thorough quantitative evaluation of surface registration algorithms in the context of scene reconstruction; our results indicate that well-known algorithms perform surprisingly poorly and that algorithms introduced in the last few years are outperformed by older approaches. Third, in addition to accuracy measurements on synthetic scenes we describe an experimental procedure for quantitative evaluation of reconstruction quality on real-world scenes in the absence of ground-truth data.

2. Related Work

The influential KinectFusion system demonstrated real-time dense reconstruction with a consumer depth camera [46], building on earlier work on range image integration [14], visual odometry [15, 48, 38], and real-time 3D reconstruction [51, 49, 45]. The original KinectFusion system used a flat voxel grid and was limited to small volumes, but this limitation has since been removed [9, 61, 47]. Alternative odometry algorithms that can improve the accuracy of the system have also been proposed [6, 61]. These systems do not detect loop closures and are limited either to compact workspaces or to fairly simple walk-along trajectories. Without dedicated loop closure handling, the complex camera paths that are necessary for comprehensive imaging of furnished indoor scenes lead to broken reconstructions [69].

A number of RGB-D reconstruction systems with integrated loop closure handling have been developed [29, 18, 28, 56, 62]. They all detect loop closures by matching individual RGB-D images using either visual features such as SIFT or SURF keypoints or through dense image registration. This approach delivers real-time performance but assumes that different images that observe the same location in the scene are sufficiently similar. It is thus liable to miss loop closures that are not certified by matching images,

as illustrated in Figure 2. Our setting is different in that real-time performance is not a requirement. High-quality off-line scene reconstruction is valuable in many application domains [19, 65, 69, 59].

Off-line global optimization for high-fidelity RGB-D reconstruction was previously considered by Zhou et al. [69, 72, 71]. Their work relied on an initialization provided by an off-the-shelf loop closure detection module [18]. It was thus prone to failure when the provided loop closure set was incomplete. Our work presents an integrated approach to loop closure detection and handling based on geometric registration and robust optimization.

Geometric registration of range data has been extensively studied [44]. A typical registration pipeline samples constellations of points on one surface and uses matching configurations on the other surface to compute candidate transformations. The challenge is that exhaustive sampling and matching are prohibitively expensive. In the past decade researchers have studied local shape descriptors that can be used for pruning and correspondence [20, 52, 55, 24], and proposed different types of constellations [2, 16, 43]. Nevertheless, misregistrations are still common in practice. Our approach uses global optimization to make the reconstruction pipeline robust to geometric registration errors.

Global optimization of range scan poses based on hypothesized pairwise relations was introduced by Lu and Milios [41] and is commonly used in robotics [13, 33, 23]. In our setting, all pairwise relations are noisy and the set of relations is heavily contaminated by outliers. Huber and Hebert [32] described an algorithm that rejects outliers by searching for a maximally consistent spanning tree in the pairwise relation graph; a similar technique has been used for reassembling fractured objects [30]. This approach assumes that the scene can be covered by accurate pairwise alignments, which is not true in our case.

Our solution is based on line processes [4]. This formulation enables effective outlier rejection using a high-performance least-squares solver. A single optimization aligns the scene and identifies the outliers even if they outnumber veridical matches. Related formulations were recently introduced in the context of robot localization [57, 58, 40, 1] and bundle adjustment [66]. In structure from motion estimation, robustness can be increased using appropriate penalty functions [12, 27, 8] or by identifying inconsistent substructures among pairwise relations between camera poses [67, 68, 50, 63]. Our work is related but focuses on dense scene reconstruction from range video. We present a dedicated formulation for dense surface reconstruction that identifies outliers by directly optimizing for surface alignment, using an objective that efficiently incorporates dense correspondence constraints. Our experiments demonstrate that the presented formulation significantly outperforms prior robust optimization frameworks that do not incorporate dense surface alignment.

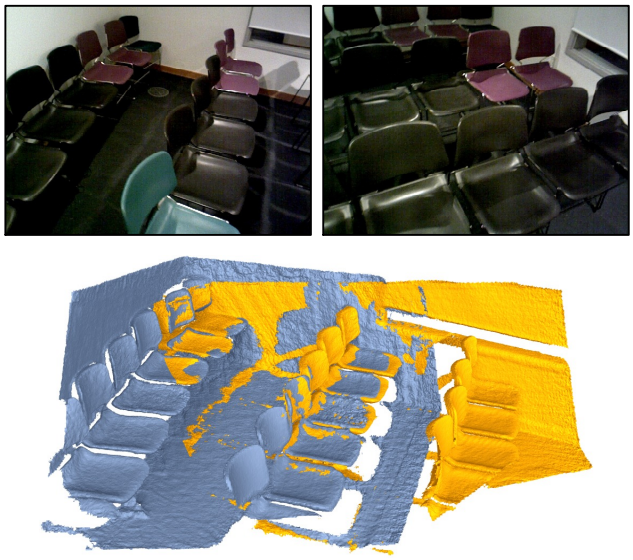


Figure 2. A challenging loop closure in the mit_32_d507 scene from the SUN3D dataset [65]. Top: this loop closure is not certified by sufficiently similar images and is missed by prior scene reconstruction pipelines. Bottom: our approach matches the underlying geometry and successfully detects the loop closure. A complete reconstruction of this scene is shown in Figure 4.

3. Overview

Fragment construction. Individual range images are noisy and incomplete. To derive more reliable information on local surface geometry, we partition the input RGB-D video into k-frame segments ($k = 50$ in all experiments), use RGB-D odometry to estimate the camera trajectory [35], and fuse the range images to obtain a surface mesh for each segment [14]. These scene fragments integrate out some of the noise in the range data and yield more reliable normal information [69, 72]. They have a larger footprint in the scene than individual images without suffering from significant odometry drift. Fragments are analogous to submaps, which are used in a number of robotic mapping systems [25, 5, 54, 10]. Let $\mathbf{P}_i = \{\mathbf{p}\}$ be the vertex set of fragment i and let $\mathbf{R}_i$ be a rigid transformation that aligns $\mathbf{P}_i$ to $\mathbf{P}_{i+1}$, computed by RGB-D odometry.

Geometric registration. Due to odometry drift, simply using the transformations $\{\mathbf{R}_i\}$ to localize the fragments yields broken reconstructions in which non-consecutive fragments that cover overlapping parts of the scene are misaligned. For this reason, we test each pair of fragments to find overlapping pairs. A geometric registration algorithm is run on each pair $(\mathbf{P}_i, \mathbf{P}_j)$. If the algorithm succeeds in aligning the fragments with sufficient overlap, a candidate

loop closure is established between fragments i and j with an associated transformation $\mathbf{T}_{ij}$.

Robust optimization. Many of the putative loop closures found by pairwise registration are false positives. We identify these spurious loop closures by optimizing a dense surface registration objective augmented by a line process over the loop closure constraints. A single least-squares objective jointly estimates the global configuration of the scene and the validity of each constraint. This formulation enables reliable pruning of erroneous constraints even when they substantially outnumber genuine loop closures.

Final model. After the final set of loop closures is identified, the odometry and loop closure transformations are refined using ICP. Pose graph optimization is used to obtain the final fragment poses $\{\mathbf{T}_i\}$ in the global frame [39]. Optional nonrigid refinement can be used to further improve the registration [71]. The registered fragments are fused into a global mesh model by volumetric integration [14].

4. Geometric Registration

We begin with a quantitative analysis of state-of-the-art surface registration algorithms on indoor scene data. This analysis motivates our approach. The analysis was performed on the augmented ICL-NUIM dataset, which augments the synthetic scenes of Handa et al. [26] with complex camera trajectories and a realistic noise model. The dataset is described in detail in supplementary material.

Given an input range video, a set of fragments $\{\mathbf{P}_i\}$ was constructed as described in Section 3. Consider a fragment pair $(\mathbf{P}_i, \mathbf{P}_j)$, with $\mathbf{P}_i$ being the smaller in terms of surface area. This pair was identified as a ground-truth loop closure if their overlap in the ground-truth scene covers more than 30% of $\mathbf{P}_i$. In this case, a ground-truth transformation $\mathbf{T}^*_{ij}$ and a set of point-to-point correspondences $\mathcal{K}^*_{ij}$ were associated with this pair.

Each algorithm was run on every fragment pair $(\mathbf{P}_i, \mathbf{P}_j)$. A computed transformation $\mathbf{T}_{ij}$ was retained as a proposed loop closure if over 30% of $\mathbf{T}_{ij}\mathbf{P}_i$ overlaps with $\mathbf{P}_j$. Each algorithm's proposed loop closures were used to measure its recall and precision. For this measurement it is not sufficient to consider the intersection of the proposed and ground-truth loop closure sets, since an algorithm may have correctly determined that there is a loop closure between $\mathbf{P}_i$ and $\mathbf{P}_j$ but produced an erroneous transformation. Therefore the candidate transformation $\mathbf{T}_{ij}$ was compared to the ground-truth transformation $\mathbf{T}^*_{ij}$. To avoid arbitrary choices in weighting different degrees of freedom in transformation space, we directly measured the effect of $\mathbf{T}_{ij}$ on the ground-truth correspondences $\mathcal{K}^*_{ij}$. A transformation is accepted if it brings these ground-truth correspondence pairs into alignment. Specifically, $\mathbf{T}_{ij}$ is considered a true positive if the RMSE of the ground-truth correspondences is below a threshold τ:

$$\frac{1}{|\mathcal{K}^*_{ij}|} \sum_{(\mathbf{p}^*, \mathbf{q}^*) \in \mathcal{K}^*_{ij}} \|\mathbf{T}_{ij}\mathbf{p}^* - \mathbf{q}^*\|^2 < \tau^2.$$

We used a fairly liberal threshold $\tau = 0.2$ meters in all experiments.

Table 1 lists the average recall and precision of different algorithms on the augmented ICL-NUIM dataset. `OpenCV` is a recent OpenCV implementation of the surface registration algorithm of Drost et al. [16]. All look-up tables were precomputed for accelerated performance. `4PCS` is the algorithm of Aiger et al. [2] and `Super 4PCS` is the recent algorithm of Mellado et al. [43]. We worked with the authors of Super 4PCS to determine the best parameter values for their approach. `PPF Integral` is our custom implementation that combines the point pair features of Drost et al. [16] with subsampling based on integral invariants [42, 20]. `PCL` is a Point Cloud Library implementation of the algorithm of Rusu et al. [52, 3]. `PCL modified` is our variant of Rusu's algorithm, described in supplementary material.

	OpenCV	4PCS	Super 4PCS	PPF Integral	PCL	PCL modified
Recall (%)	5.3	20.0	17.8	32.5	44.9	**59.2**
Precision (%)	1.6	8.9	10.4	7.1	14.0	**19.6**
Runtime (sec)	10	380	62	83	**3**	8

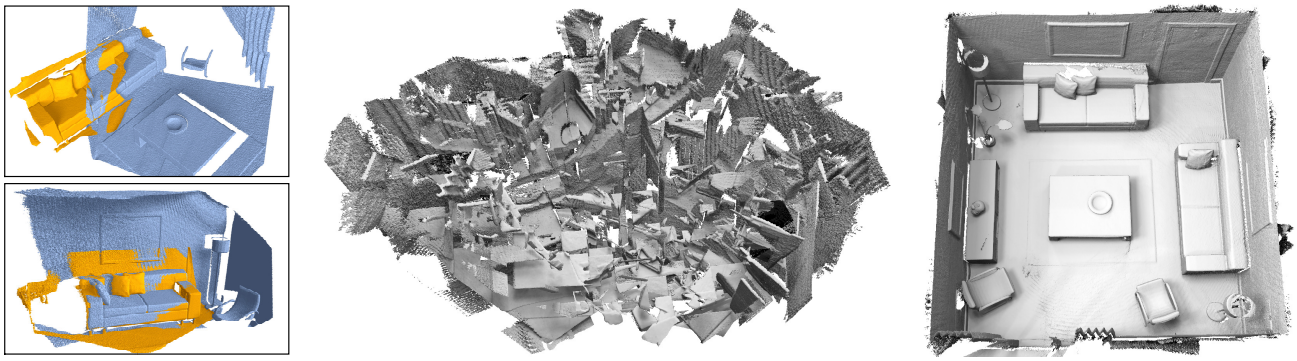
Table 1. Performance of geometric registration algorithms. Average running time for aligning two fragments was measured using a single thread on an Intel Core i7-3770 CPU clocked at 3.5 GHz.

Surprisingly, the algorithm of Rusu et al. outperforms all other approaches, including more recent ones. Based on this experiment, our pipeline uses the `PCL modified` algorithm for pairwise geometric registration.

As the results in Table 1 indicate, the precision of even the highest-performing geometric registration algorithms is below 20%. We attribute this primarily to the limited discriminative power of surface geometry that was sampled at limited range, resolution, and field of view, and corrupted by noise and distortion. This aliasing permits reasonable recall but limits precision. As illustrated in Figure 3(a), some false positive alignments are very plausible when considered independently. Thus, rather than attempt to develop a pairwise surface registration procedure with high recall and near-perfect precision, we show in Section 5 that these characteristics can be achieved by a global analysis of the scene.

5. Robust Optimization

The analysis in Section 4 indicates that most loop closures identified by pairwise surface matching are false positives. We now show that global optimization can be used to achieve near-perfect loop closure precision with almost no decrease in recall.



(a) Erroneous alignments (b) Optimization with all pairwise alignments (c) Optimization with line processes

Figure 3. The precision problem, illustrated on the Living room 1 sequence from the augmented ICL-NUIM dataset. (a) False positive alignments of fragment pairs. Note that the alignments look plausible. (b) Pose graph optimization with all pairwise alignments identified by geometric registration. (c) Optimization with line processes.

Consider a pose graph with vertices $\{\mathbf{P}_i\}$ and edges $\{\mathbf{R}_i\} \cup \{\mathbf{T}_{ij}\}$ [23]. Our goal is to compute a set of poses $\mathbb{T} = \{\mathbf{T}_i\}$ that localizes the fragments in the global coordinate frame. This can be expressed as an objective of the form

$$E(\mathbb{T}) = \sum_i f(\mathbf{T}_i, \mathbf{T}_{i+1}, \mathbf{R}_i) + \sum_{i,j} f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{T}_{ij}). \quad (1)$$

The challenge is that most of the transformations $\mathbf{T}_{ij}$ are incorrect and will corrupt the optimized configuration, as shown in Figure 3(b). We thus add a line process $\mathbb{L} = \{l_{ij}\}$ over the putative loop closures. The variable l_{ij} ranges over $[0, 1]$ and models the validity of the corresponding loop closure. $\mathbb{L}$ and $\mathbb{T}$ are optimized jointly:

$$\begin{aligned} E(\mathbb{T}, \mathbb{L}) &= \sum_i f(\mathbf{T}_i, \mathbf{T}_{i+1}, \mathbf{R}_i) \\ &+ \sum_{i,j} l_{ij} f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{T}_{ij}) \\ &+ \mu \sum_{i,j} \Psi(l_{ij}). \end{aligned} \quad (2)$$

The prior term $\Psi(l_{ij})$ expresses a belief that proposed loop closures are genuine: $\Psi(l_{ij}) = (\sqrt{l_{ij}} - 1)^2$. Intuitively, this term aims to maximize the number of selected loop closures ($l_{ij} \to 1$). However, if a constraint distorts the configuration and causes a disproportionate increase in the alignment terms it can be smoothly disabled ($l_{ij} \to 0$).

An alignment term $f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X})$ measures the inconsistency between poses $\mathbf{T}_i$ and $\mathbf{T}_j$ and relative pose $\mathbf{X}$. We define this function in terms of dense surface alignment. Let $\mathcal{K}_{ij}$ be the set of correspondence pairs between $\mathbf{X}\mathbf{P}_i$ and $\mathbf{P}_j$ that are within distance $\varepsilon = 0.05$ m. (ε was set based on typical sensor noise magnitudes [36].) Define $f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X})$ as the sum of squared distances between corresponding points in $\mathbf{T}_i\mathbf{P}_i$ and $\mathbf{T}_j\mathbf{P}_j$:

$$\begin{aligned} f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X}) &= \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{K}_{ij}} \|\mathbf{T}_i\mathbf{p} - \mathbf{T}_j\mathbf{q}\|^2 & (3) \\ &\approx \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{K}_{ij}} \|\mathbf{T}_i\mathbf{p} - \mathbf{T}_j\mathbf{X}\mathbf{p}\|^2 & (4) \\ &= \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{K}_{ij}} \|\mathbf{X}^{-1}\mathbf{T}_j^{-1}\mathbf{T}_i\mathbf{p} - \mathbf{p}\|^2. & (5) \end{aligned}$$

Line (4) uses the proximity of correspondence pairs, which is guaranteed by construction: $(\mathbf{p}, \mathbf{q}) \in \mathcal{K}_{ij} \Rightarrow \|\mathbf{X}\mathbf{p} - \mathbf{q}\| < \varepsilon$.

Use a standard local parameterization to represent $\mathbf{X}^{-1}\mathbf{T}_j^{-1}\mathbf{T}_i$ as a 6-vector $\xi = (\omega, \mathbf{t}) = (\alpha, \beta, \gamma, a, b, c)$, which collects a rotational component ω and a translation $\mathbf{t}$. Locally, when $\mathbf{T}_j^{-1}\mathbf{T}_i \approx \mathbf{X}$,

$$\mathbf{X}^{-1}\mathbf{T}_j^{-1}\mathbf{T}_i \approx \begin{pmatrix} 1 & -\gamma & \beta & a \\ \gamma & 1 & -\alpha & b \\ -\beta & \alpha & 1 & c \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (6)$$

Thus

$$\mathbf{X}^{-1}\mathbf{T}_j^{-1}\mathbf{T}_i\mathbf{p} \approx \mathbf{p} + \omega \times \mathbf{p} + \mathbf{t}.$$

Equation (5) can be locally approximated as

$$\begin{aligned} f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X}) &\approx \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{K}_{ij}} \|\omega \times \mathbf{p} + \mathbf{t}\|^2 \\ &= \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{K}_{ij}} \left\| \left[-[\mathbf{p}]_\times \mid \mathbf{I} \right] \xi \right\|^2, \end{aligned} \quad (7)$$

where $[\mathbf{p}]_\times$ is the skew-symmetric matrix form of the cross product with $\mathbf{p}$, and $\mathbf{I}$ is the 3×3 identity matrix. Define $\mathbf{G}_\mathbf{p} = \left[-[\mathbf{p}]_\times \mid \mathbf{I} \right]$.

$$
\begin{aligned}
f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X}) &\approx \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}_{ij}} \|\mathbf{G}_\mathbf{p}\xi\|^2 \\
&= \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}_{ij}} \xi^\top \mathbf{G}_\mathbf{p}^\top \mathbf{G}_\mathbf{p}\xi \\
&= \xi^\top \left(\sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}_{ij}} \mathbf{G}_\mathbf{p}^\top \mathbf{G}_\mathbf{p} \right) \xi. \quad (8)
\end{aligned}
$$

Since $\mathbf{G_p}$ is constant, $f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X})$ can be approximated by the quadratic form $\xi^\top \Lambda \xi$, where the covariance

$$
\Lambda = \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}_{ij}} \mathbf{G}_\mathbf{p}^\top \mathbf{G}_\mathbf{p} \quad (9)
$$

need only be computed once for each alignment term.

The parameter μ balances the strength of the prior term and the alignment terms. Given the above derivation of the alignment terms, μ is defined to be proportional to the average cardinality of the correspondence sets $\mathcal{K}_{ij}$, denoted by κ: $\mu = \tau^2\kappa$. τ is the distance threshold used in Section 4 and has the same underlying semantics. Intuitively, when an error $f(\mathbf{T}_i, \mathbf{T}_j, \mathbf{X})$ exceeds μ, it outweighs the corresponding prior term.

The objective (2) is optimized using $\mathbf{g^2o}$ [39] and loop closures with $l_{ij} < 0.25$ are pruned. The remaining loop closures are used to construct the final model as described in Section 3.

This formulation is extremely effective. Table 2 summarizes the effect of the presented formulation on the augmented ICL-NUIM dataset. The optimization increases the average precision of the loop closure set by a factor of five, from below 20% to 97.7%. The average recall decreases by only 1.4%.

	Before pruning		After pruning	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Living room 1	61.2	27.2	57.6	95.1
Living room 2	49.7	17.0	49.7	97.4
Office 1	64.4	19.2	63.3	98.3
Office 2	61.5	14.9	60.7	100.0
Average	59.2	19.6	57.8	97.7

Table 2. The effect of robust optimization. The optimization increases the average precision of the loop closure set from 19.6% to 97.7%.

The basic formulation (2) is an application of line processes [4] and has been used for pose graph optimization before [57]. Our work differs by incorporating surface alignment into the objective. To evaluate the impact of this formulation, we measured the loop closure precision achieved by the basic switchable constraints approach of Sünderhauf and Protzel (SC) [57], the expectation maximization algorithm of Lee et al. (EM) [40], and our formulation. The results on the augmented ICL-NUIM dataset are reported in Table 3. The prior approaches do improve the precision of the loop closure set, but the improvement is not sufficient for satisfactory reconstruction, as shown in Section 6.2. Our formulation achieves near-perfect precision and significantly improves reconstruction accuracy.

	Original	SC [57]	EM [40]	Ours
Living room 1	27.2	54.6	39.6	**95.1**
Living room 2	17.0	23.5	20.5	**97.4**
Office 1	19.2	39.6	33.7	**98.3**
Office 2	14.9	25.2	19.7	**100.0**
Average	19.6	35.7	28.4	**97.7**

Table 3. The effect of surface alignment modeling. From left to right: precision of the loop closure set without pruning, optimization using basic switchable constraints [58], optimization using expectation maximization [40], and optimization using our formulation.

6. Evaluation

6.1. Datasets

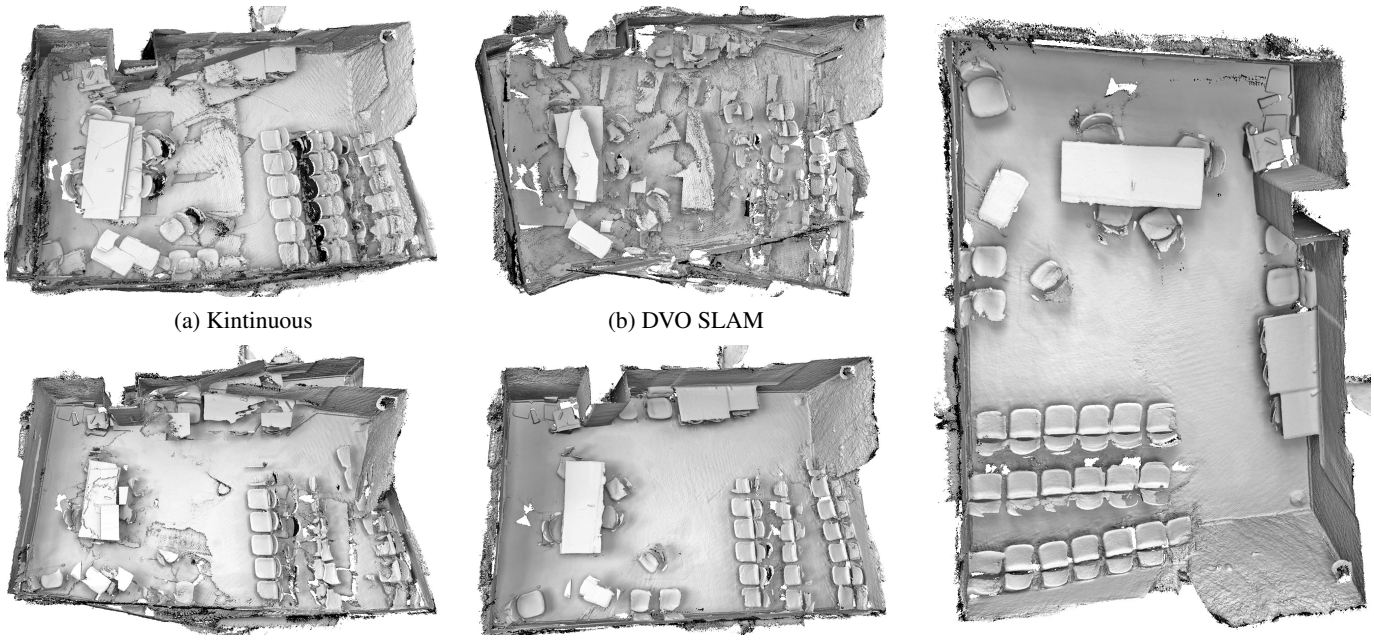
Augmented ICL-NUIM dataset. Our first dataset is based on the synthetic environments provided by Handa et al. [26]. The availability of ground-truth surface geometry enables precise measurement of reconstruction accuracy. The dataset includes two models of indoor environments: a living room and an office. We have augmented the dataset in a number of ways to adapt it for evaluation of complete scene reconstruction pipelines. We have verified with the authors that these extensions are in line with the intended usage of the dataset. Our experiments are conducted on four input sequences that model thorough handheld imaging for the purpose of comprehensive reconstruction: Living room 1, Living room 2, Office 1, and Office 2. The augmented dataset is described in detail in supplementary material.

SUN3D dataset. Our second dataset is based on the SUN3D database of indoor scenes [65]. The original dataset released by Xiao et al. includes an off-line system for automatic scene reconstruction based on bundle adjustment, which we use for comparison. It also provides a number of reconstructions that were created with manual assistance, using an interactive interface that lets the user establish object-level correspondences across the input video. Xiao et al. provided models of eight scenes reconstructed with such manual assistance. We focus on these scenes, since the manually-assisted reconstructions are a useful reference.

Running time. Running times for all steps of our pipeline are reported in supplementary material.

6.2. Synthetic scenes

To evaluate surface reconstruction accuracy on ICL-NUIM scenes we use the error measures proposed by Handa et al., specifically the mean and median of the distances of the reconstructed surfaces to the ground-truth surfaces. For each sequence we evaluate four reconstruction pipelines: Kintinuous [61], DVO SLAM [34], the automatic bundle



(a) Kintinuous (b) DVO SLAM (c) SUN3D SfM (d) Our result (e) Optional non-rigid refinement

Figure 4. Reconstruction of the mit_32_d507 scene from the SUN3D dataset. (a) Reconstruction produced by Kintinuous [61]. (b) Reconstruction produced by DVO SLAM [34]. (c) Reconstruction produced by the off-line RGB-D structure-from-motion pipeline of Xiao et al. [65]. (d) Reconstruction produced by our approach. (e) An optional non-rigid refinement of our result using SLAC [71].

adjustment pipeline provided by Xiao et al. [65] (SUN3D SfM), and our approach. Qualitative results are shown in supplementary material. For reference, we also measure the accuracy of a model obtained by fusing the input depth images along the ground-truth trajectory (GT trajectory); these depth images are affected by the simulated sensor noise and the reconstructed model is thus imperfect. Mean distances are reported in Table 4, median distances in supplementary material. The presented approach considerably outperforms the other reconstruction pipelines. The average mean error is reduced by a factor of 2 relative to the closest alternative approach (SUN3D SfM). The average median error is reduced by a factor of 2.7. Note that this is a direct evaluation of the metric accuracy of reconstructed models.

	Kintinuous	DVO SLAM	SUN3D SfM	Ours	GT trajectory
Living room 1	0.22	0.21	0.09	**0.04**	0.04
Living room 2	0.14	0.06	0.07	**0.07**	0.04
Office 1	0.13	0.11	0.13	**0.03**	0.03
Office 2	0.13	0.10	0.09	**0.04**	0.03
Average	0.16	0.12	0.10	**0.05**	0.04

Table 4. Reconstruction accuracy on ICL-NUIM sequences. Mean distance of each reconstructed model to the ground-truth surface (in meters). Our approach reduces the average error by a factor of 2 relative to the closest alternative approach.

For completeness, we have also measured the accuracy of the estimated camera trajectories using the RMSE metric described by Handa et al. The results are reported in supplementary material. Trajectories estimated by our approach are considerably more accurate, with average RMSE reduced by a factor of 2.2 relative to the closest alternative approach. Note that trajectory accuracy is only an indirect measure of reconstruction accuracy: the metric surface accuracy measurements reported in Table 4 are more informative.

We have also conducted a controlled evaluation of the effects of different components of our pipeline on final reconstruction accuracy. Specifically, we substituted the geometric loop closure detection pipeline presented in Section 4 with the state-of-the-art image-based pipeline of Kerl et al. [34]. (For these experiments, all settings for image-based loop closure [34] were set to maximize accuracy, and loop closure detection was performed between every single pair of frames.) Independently, we substituted the robust optimization formulation presented in Section 5 with basic switchable constraints [57] or expectation maximization [40]. (These algorithms were also considered in Section 5.) The results are reported in Table 5. The presented pipeline yields much higher reconstruction accuracy.

6.3. Real-world scenes

Experimental procedure. Quantitative evaluation on real-world scenes is challenging because there is no ground-truth surface model. We have therefore developed and validated a perceptual evaluation procedure. Extensive pairwise comparisons were collected for all pairs of recon-

	SC [57]		EM [40]		Ours
	[34]	geometric	[34]	geometric	
Living room 1	0.25	0.32	0.46	0.66	**0.04**
Living room 2	0.26	0.40	0.26	0.65	**0.07**
Office 1	0.11	0.36	0.22	0.56	**0.03**
Office 2	0.52	0.27	0.56	0.48	**0.04**
Average	0.28	0.34	0.35	0.59	**0.05**

Table 5. Controlled evaluation of different components of the presented pipeline. Reconstruction accuracy on ICL-NUIM sequences: mean distances to ground-truth models, in meters. Replacing our robust optimization formulation with basic switchable constraints (SC) or expectation maximization (EM) results in significant degradation of reconstruction accuracy.

structed models for each input sequence. Experiments were conducted using Amazon Mechanical Turk. The pairwise comparison interface and the experimental protocol are described and demonstrated in detail in supplementary material. The collected pairwise comparisons were used to compute a numerical score for each reconstructed model via Balanced Rank Estimation (BRE) [60]. The BRE scores are in the range $[-1, 1]$, higher is better.

Validation. The experimental procedure was used to collect pairwise comparisons and compute BRE scores for the ICL-NUIM sequences. We evaluated models reconstructed by Kintinuous, DVO SLAM, SUN3D SfM, and our approach. For reference, we also evaluated models produced by integration of the noisy input data along the ground-truth trajectory. 8,960 pairwise comparisons were collected. The resulting BRE scores are shown in Table 6. The order of the average BRE scores is identical to the order of the average mean ground-truth distances reported in Table 4. Note that the BRE scores are not linearly related to the ground-truth distance measures, nor can they be since the distance measures are in the range $[0, \infty)$ and the BRE scores are in the range $[-1, 1]$.

	Kintinuous	DVO SLAM	SUN3D SfM	Ours	GT trajectory
Living room 1	-0.53	-0.90	0.02	**0.47**	0.94
Living room 2	-0.89	-0.65	-0.13	**0.66**	0.89
Office 1	-0.71	-0.41	-0.15	**0.09**	0.98
Office 2	-0.83	-0.57	-0.11	**0.58**	0.90
Average	-0.74	-0.63	-0.09	**0.45**	0.93

Table 6. Perceptual evaluation on ICL-NUIM sequences. BRE scores computed from pairwise comparisons performed on Amazon Mechanical Turk. The order of the average BRE scores is identical to the order of the average mean and median ground-truth distances reported in Table 4 and in supplementary material.

Experimental results. The same experimental procedure was applied to the eight SUN3D sequences. We evaluated models reconstructed by Kintinuous, DVO SLAM, SUN3D SfM, and our approach. For reference, we also included the manually-assisted reconstructions provided by Xiao et al. 17,640 pairwise comparisons were collected. The resulting BRE scores are shown in Table 7. The presented approach outperforms all other automatic reconstruction pipelines. It is also ranked more highly than the manually-assisted reconstructions on 6 out of 8 sequences. We ascribe this to limitations of the SUN3D interactive labeling interface, which focuses on object labeling and establishes only region-level correspondences. Reconstructed models for one of the scenes are shown in Figure 4.

	DVO SLAM	Kintinuous	SUN3D SfM	Ours	SUN3D manual
hotel_umd	-0.61	-0.45	-0.02	**0.66**	0.56
harvard_c5	-0.49	-0.01	-0.65	**0.94**	0.11
harvard_c6	-0.97	0.05	-0.01	**0.96**	-0.15
harvard_c8	-0.70	-0.61	0.39	**0.65**	0.46
mit_32_d507	-0.78	-0.28	-0.02	**0.74**	0.36
mit_76_studyroom	-0.52	-0.47	0.35	**0.50**	0.19
mit_dorm_next_sj	-0.26	-0.20	-0.23	**0.10**	0.65
mit_lab_hj	-0.12	-0.57	0.03	**0.22**	0.50
Average	-0.56	-0.32	-0.02	**0.60**	0.33

Table 7. Perceptual evaluation on SUN3D scenes. BRE scores computed from pairwise comparisons performed on Amazon Mechanical Turk. The presented approach outperforms all other automatic reconstruction pipelines.

7. Conclusion

We presented an approach to scene reconstruction from RGB-D video. The key idea is to combine geometric registration with global optimization based on line processes. The optimization makes the pipeline robust to erroneous geometric alignments, which are unavoidable due to aliasing in the input. Experimental results demonstrate that the presented approach significantly improves the fidelity of indoor scene models produced from consumer-grade video.

The presented pipeline is not foolproof. First, if the input video does not contain loop closures that indicate global geometric relations, odometry drift can accumulate and distort the reconstructed model. Real-time feedback that guides the operator to close loops would help. We believe that the presented approach can be adapted for real-time operation, which would assist the acquisition of complete scene models. Second, our pipeline does not take into account the possibility of catastrophic odometry failure, which would result in missing or misshapen fragments. This could be addressed by modeling uncertainty not only at the inter-fragment level but also in the individual fragment shapes. Integration of inertial data would also be useful in challenging scenarios.

Acknowledgements

We thank Andreas Geiger for helpful discussions, the authors of the ICL-NUIM and SUN3D datasets for their data and relevant discussions, and the authors of Super 4PCS for testing their algorithm on our data.

References

[1] P. Agarwal, G. Grisetti, G. D. Tipaldi, L. Spinello, W. Burgard, and C. Stachniss. Experimental analysis of dynamic covariance scaling for robust map optimization under bad initial estimates. In *ICRA*, 2014. 3

[2] D. Aiger, N. J. Mitra, and D. Cohen-Or. 4-points congruent sets for robust pairwise surface registration. *ACM Transactions on Graphics*, 27(3), 2008. 3, 4

[3] A. Aldoma, Z. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze. Point cloud library: Three-dimensional object recognition and 6 DoF pose estimation. *IEEE Robotics and Automation Magazine*, 19(3), 2012. 4

[4] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *IJCV*, 19(1), 1996. 1, 3, 6

[5] M. Bosse, P. M. Newman, J. J. Leonard, and S. J. Teller. Simultaneous localization and map building in large-scale cyclic environments using the Atlas framework. *International Journal of Robotics Research*, 23(12), 2004. 3

[6] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers. Real-time camera tracking and 3D reconstruction using signed distance functions. In *RSS*, 2013. 2

[7] R. Cabral and Y. Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *CVPR*, 2014. 1

[8] A. Chatterjee and V. M. Govindu. Efficient and robust large-scale rotation averaging. In *ICCV*, 2013. 3

[9] J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics*, 32(4), 2013. 2

[10] L. A. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardós. Mapping large loops with a single hand-held camera. In *RSS*, 2007. 3

[11] N. Corso and A. Zakhor. Indoor localization algorithms for an ambulatory human operated 3D mobile mapping system. *Remote Sensing*, 5(12), 2013. 1

[12] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *PAMI*, 35(12), 2013. 3

[13] M. J. Cummins and P. M. Newman. FAB-MAP: probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6), 2008. 3

[14] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996. 2, 3, 4

[15] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, 2003. 2

[16] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *CVPR*, 2010. 3, 4

[17] H. Du, P. Henry, X. Ren, M. Cheng, D. B. Goldman, S. M. Seitz, and D. Fox. Interactive 3D modeling of indoor environments with a consumer depth camera. In *UbiComp*, 2011. 1

[18] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard. 3-D mapping with an RGB-D camera. *IEEE Transactions on Robotics*, 30(1), 2014. 1, 2, 3

[19] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, 2009. 1, 3

[20] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann. Robust global registration. In *Symposium on Geometry Processing*, 2005. 3, 4

[21] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *PAMI*, 14(3), 1992. 1

[22] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *PAMI*, 6(6), 1984. 1

[23] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2(4), 2010. 3, 5

[24] Y. Guo, F. A. Sohel, M. Bennamoun, M. Lu, and J. Wan. Rotational projection statistics for 3D local surface description and object recognition. *IJCV*, 105(1), 2013. 3

[25] J. Gutmann and K. Konolige. Incremental mapping of large cyclic environments. In *CIRA*, 1999. 3

[26] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *ICRA*, 2014. 1, 4, 6

[27] R. I. Hartley, K. Aftab, and J. Trumpf. L1 rotation averaging using the Weiszfeld algorithm. In *CVPR*, 2011. 3

[28] P. Henry, D. Fox, A. Bhowmik, and R. Mongia. Patch volumes: Segmentation-based consistent mapping with RGB-D cameras. In *3DV*, 2013. 2

[29] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research*, 31(5), 2012. 1, 2

[30] Q. Huang, S. Flöry, N. Gelfand, M. Hofer, and H. Pottmann. Reassembling fractured objects by geometric matching. *ACM Transactions on Graphics*, 25(3), 2006. 3

[31] D. F. Huber, O. T. Carmichael, and M. Hebert. 3-D map reconstruction from range data. In *ICRA*, 2000. 1

[32] D. F. Huber and M. Hebert. Fully automatic registration of multiple 3D data sets. *Image and Vision Computing*, 21(7), 2003. 3

[33] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental smoothing and mapping. *IEEE Transactions on Robotics*, 24(6), 2008. 3

[34] C. Kerl, J. Sturm, and D. Cremers. Dense visual SLAM for RGB-D cameras. In *IROS*, 2013. 6, 7, 8

[35] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *ICRA*, 2013. 3

[36] K. Khoshelham and S. O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12(2), 2012. 5

[37] Y. M. Kim, J. Dolson, M. Sokolsky, V. Koltun, and S. Thrun. Interactive acquisition of residential floor plans. In *ICRA*, 2012. 1

[38] G. Klein and D. W. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007. 2

[39] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G^2o: A general framework for graph optimization. In *ICRA*, 2011. 4, 6

[40] G. H. Lee, F. Fraundorfer, and M. Pollefeys. Robust pose-graph loop-closures with expectation-maximization. In *IROS*, 2013. 3, 6, 7, 8

[41] F. Lu and E. E. Milios. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, 4(4), 1997. 3

[42] S. Manay, D. Cremers, B. Hong, A. J. Yezzi, and S. Soatto. Integral invariants for shape matching. *PAMI*, 28(10), 2006. 4

[43] N. Mellado, D. Aiger, and N. J. Mitra. Super 4PCS: Fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33(5), 2014. 3, 4

[44] A. S. Mian, M. Bennamoun, and R. A. Owens. Automatic correspondence for 3D modeling: an extensive review. *International Journal of Shape Modeling*, 11(2), 2005. 3

[45] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *CVPR*, 2010. 2

[46] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 1, 2

[47] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics*, 32(6), 2013. 2

[48] D. Nistér, O. Naroditsky, and J. R. Bergen. Visual odometry. In *CVPR*, 2004. 2

[49] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. N. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *IJCV*, 78(2-3), 2008. 2

[50] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *CVPR*, 2011. 3

[51] S. Rusinkiewicz, O. A. Hall-Holt, and M. Levoy. Real-time 3D model acquisition. *ACM Transactions on Graphics*, 21(3), 2002. 2

[52] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, 2009. 3, 4

[53] A. Sankar and S. M. Seitz. Capturing indoor scenes with smartphones. In *UIST*, 2012. 1

[54] S. Se, D. G. Lowe, and J. J. Little. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21(3), 2005. 3

[55] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard. Point feature extraction on 3D range scans taking into account object boundaries. In *ICRA*, 2011. 3

[56] F. Steinbrücker, C. Kerl, and D. Cremers. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In *ICCV*, 2013. 2

[57] N. Sünderhauf and P. Protzel. Switchable constraints for robust pose graph SLAM. In *IROS*, 2012. 3, 6, 7, 8

[58] N. Sünderhauf and P. Protzel. Switchable constraints vs. max-mixture models vs. RRR – a comparison of three approaches to robust pose graph SLAM. In *ICRA*, 2013. 3, 6

[59] E. Turner, P. Cheng, and A. Zakhor. Fast, automated, scalable generation of textured 3D models of indoor environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(3), 2015. 1, 3

[60] F. Wauthier, M. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *ICML*, 2013. 8

[61] T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and J. McDonald. Robust real-time visual odometry for dense RGB-D mapping. In *ICRA*, 2013. 2, 6, 7

[62] T. Whelan, J. McDonald, M. Kaess, and J. J. Leonard. Deformation-based loop closure for large scale dense RGB-D SLAM. In *IROS*, 2013. 1, 2

[63] K. Wilson and N. Snavely. Network principles for SfM: Disambiguating repeated structures with local context. In *ICCV*, 2013. 3

[64] J. Xiao and Y. Furukawa. Reconstructing the world's museums. In *ECCV*, 2012. 1

[65] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, 2013. 1, 3, 6, 7

[66] C. Zach. Robust bundle adjustment revisited. In *ECCV*, 2014. 3

[67] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3D structure and motion? In *CVPR*, 2008. 3

[68] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, 2010. 3

[69] Q.-Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics*, 32(4), 2013. 1, 2, 3

[70] Q.-Y. Zhou and V. Koltun. Color map optimization for 3D reconstruction with consumer depth cameras. *ACM Transactions on Graphics*, 33(4), 2014. 1

[71] Q.-Y. Zhou and V. Koltun. Simultaneous localization and calibration: Self-calibration of consumer depth cameras. In *CVPR*, 2014. 3, 4, 7

[72] Q.-Y. Zhou, S. Miller, and V. Koltun. Elastic fragments for dense scene reconstruction. In *ICCV*, 2013. 3