Robust Reconstruction of Indoor Scenes
Supplementary Material

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

Appendix A: Qualitative results

Reconstructed models for two scenes from the SUN3D dataset are shown in Figures 1 and 2.

Appendix B: Pairwise registration algorithm

This appendix summarizes the pairwise registration algorithm used in our pipeline (`PCL modified`). This supports Section 4 in the paper. Given a pair of fragments $(\mathbf{P}_i, \mathbf{P}_j)$, the algorithm computes a rigid transformation $\mathbf{T}_{ij}$ that aligns them as well as possible. The algorithm is based on the work of Rusu et al. [7], with two main modifications. First, we sample constellations of four points rather than three, which slightly lowers recall but significantly increases precision. Second, we have added a number of validation steps that rapidly prune poorly matching constellations and allow a larger number of constellations to be tested within the same computation budget, yielding higher recall.

The steps are given in Algorithm 1. The two fragments are uniformly covered by sample points. For each sample point $\mathbf{s}$, we compute its FPFH descriptor $\mathbf{F}(\mathbf{s})$, a 33-dimensional descriptor of local shape. The descriptors are used to match samples across fragments. In particular, given a sample point $\mathbf{p} \in \mathbf{P}_i$, it is matched to a sample $\mathbf{q}_\mathbf{p} \in \mathbf{P}_j$ that is closest in descriptor space:

$$\mathbf{q}_\mathbf{p} = \underset{\mathbf{q} \in \mathbf{P}_j}{\arg\min} \|\mathbf{F}(\mathbf{p}) - \mathbf{F}(\mathbf{q})\|^2. \quad (1)$$

A transformation $\mathbf{T}_{ij}$ is then computed using RANSAC. Each iteration randomly samples four pairs $(\mathbf{p}, \mathbf{q}_\mathbf{p})$ and computes a transformation $\mathbf{T}$ that aligns the four samples from $\mathbf{P}_i$ to their counterparts in $\mathbf{P}_j$ in a least-squares sense. This transformation is passed through a number of validation steps. If multiple transformations pass the validation, the algorithm outputs the transformation that maximizes the overlap of the registered fragments.

Algorithm 1: Pairwise registration

input : A pair of fragments $(\mathbf{P}_i, \mathbf{P}_j)$
output : Transformation $\mathbf{T}_{ij}$ and correspondence set $\mathcal{K}_{ij}$

Downsample $\mathbf{P}_i = \{\mathbf{p}\}$ and $\mathbf{P}_j = \{\mathbf{q}\}$;
Compute normals $\{\mathbf{n}_\mathbf{p}\}$ and $\{\mathbf{n}_\mathbf{q}\}$;
Compute FPFH features $\{\mathbf{F}(\mathbf{p})\}$ and $\{\mathbf{F}(\mathbf{q})\}$;
$\mathbf{T}_{ij} \leftarrow \emptyset$, $\mathcal{K}_{ij} \leftarrow \emptyset$;
max_correspondences $\leftarrow 0$;
for $i \leftarrow 1$ **to** *max_iteration* **do**
  // RANSAC iteration;
  Randomly pick four points $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$ from $\mathbf{P}_i$;
  Find matching samples $(\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3)$ on $\mathbf{P}_j$ using equation (1);
  Compute transformation $\mathbf{T}$ that aligns these two sets of samples;
  // Validation;
  if $\angle(\mathbf{T}\mathbf{n}_{\mathbf{p}_k}, \mathbf{n}_{\mathbf{q}_k}) > 30^\circ$ **then**
    continue;
  if $\|\mathbf{p}_k - \mathbf{p}_{k+1}\| < 0.9\,\|\mathbf{q}_k - \mathbf{q}_{k+1}\|$ or vice versa **then**
    continue;
  Compute correspondences $\mathcal{K}$ between $\mathbf{T}\mathbf{P}_i$ and $\mathbf{P}_j$;
  if $|\mathcal{K}| < \frac{1}{3}\min(|\mathbf{P}_i|, |\mathbf{P}_j|)$ **then**
    continue;
  // Update;
  if $|\mathcal{K}| >$ *max_correspondences* **then**
    $\mathbf{T}_{ij} \leftarrow \mathbf{T}$, $\mathcal{K}_{ij} \leftarrow \mathcal{K}$;
    max_correspondences $\leftarrow |\mathcal{K}|$;

Appendix C: Running times

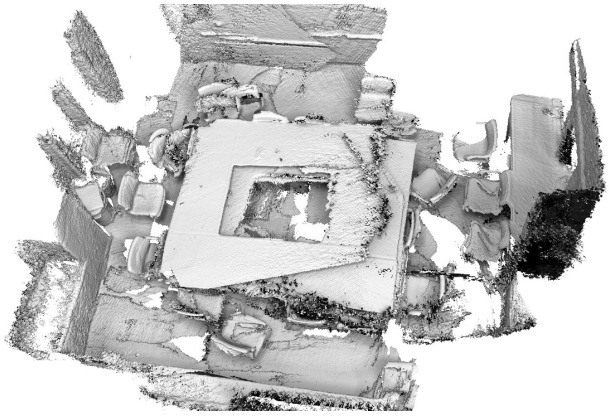
Table 1 reports running times for all steps of our pipeline. This supports Section 6.1 in the paper.

Appendix D: Augmented ICL-NUIM dataset

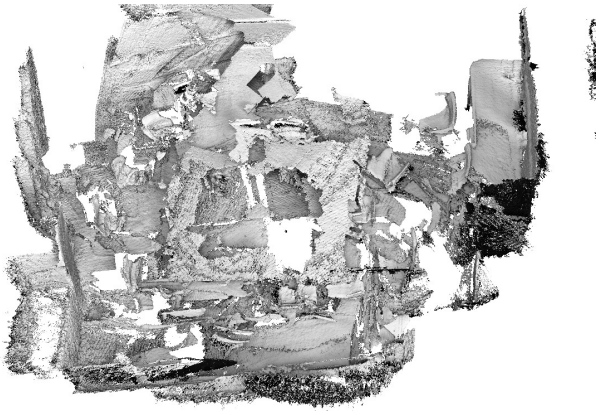
This appendix describes the augmented ICL-NUIM dataset. This supports Section 6.1 in the paper.

The dataset is based on the synthetic environments provided by Handa et al. [2]. The authors provided two models of indoor environments – a living room and an office –

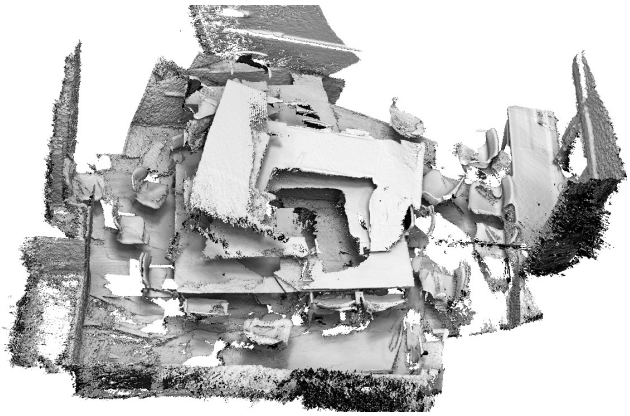
[*]Joint first authors
[†]Stanford University
[‡]Intel Labs



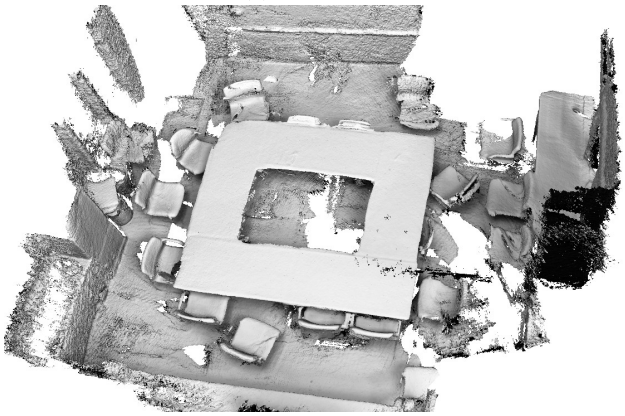
(a) Kintinuous



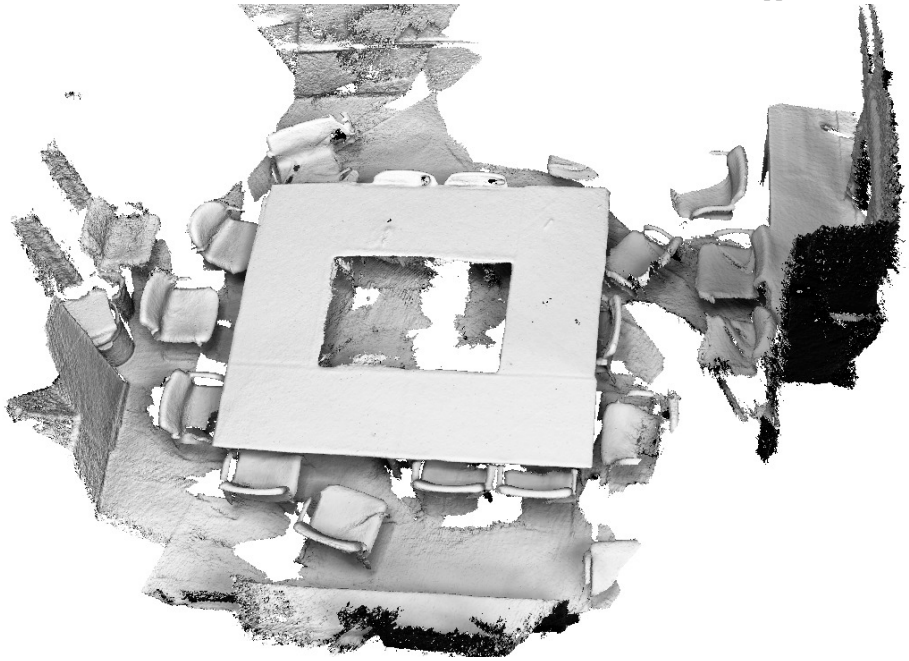
(b) DVO SLAM



(c) SUN3D SfM

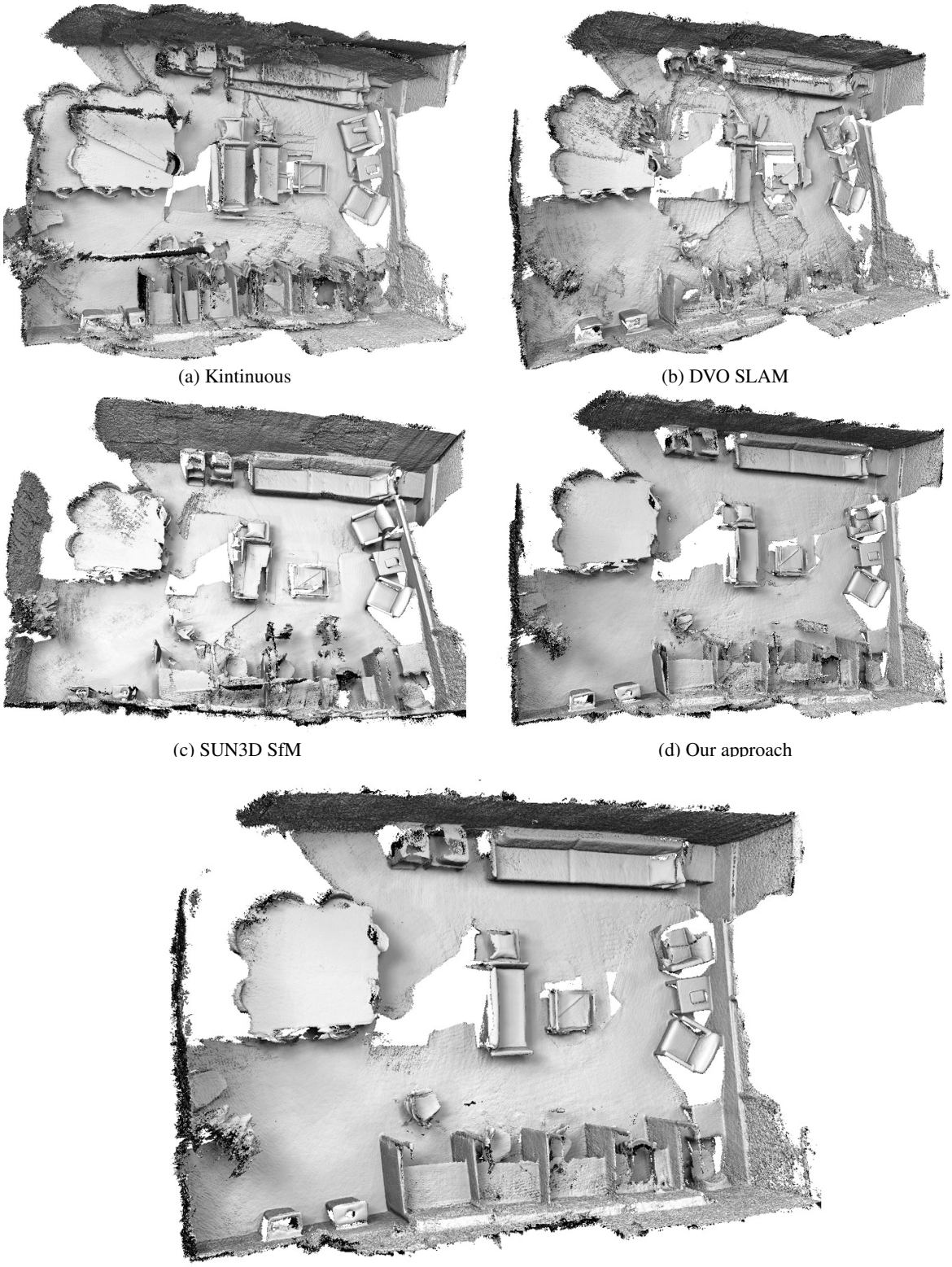


(d) Our approach



(e) Optional non-rigid refinement

Figure 1. Reconstruction of the harvard_c6 scene from the SUN3D dataset. (a) Reconstruction produced by Kintinuous [9]. (b) Reconstruction produced by DVO SLAM [4]. (c) Reconstruction produced by the off-line RGB-D structure-from-motion pipeline of Xiao et al. [10]. (d) Reconstruction produced by our approach. (e) An optional non-rigid refinement of our result using SLAC [11].



(a) Kintinuous

(b) DVO SLAM

(c) SUN3D SfM

(d) Our approach

(e) Optional non-rigid refinement

Figure 2. Reconstruction of the mit_76_studyroom scene from the SUN3D dataset. (a) Reconstruction produced by Kintinuous [9]. (b) Reconstruction produced by DVO SLAM [4]. (c) Reconstruction produced by the off-line RGB-D structure-from-motion pipeline of Xiao et al. [10]. (d) Reconstruction produced by our approach. (e) An optional non-rigid refinement of our result using SLAC [11].

Sequence		# of frames	Fragment creation	Geometric registration	Robust optimization	ICP refinement	Integration	Total time	# of triangles
Apartment		17,391	10	75	<1	194	107	387	15,759,593
Synthetic	Living room 1	2,870	2	29	<1	120	26	178	12,380,759
	Living room 2	2,350	2	24	<1	32	5	64	8,042,303
	Office 1	2,690	2	29	<1	46	8	86	12,989,830
	Office 2	2,538	2	27	<1	41	10	81	7,613,056
SUN3D	hotel_umd	1,869	1	15	<1	11	5	33	9,060,783
	harvard_c5	2,063	1	18	<1	50	5	75	4,913,239
	harvard_c6	1,517	1	10	<1	13	4	29	6,143,208
	harvard_c8	1,003	1	4	<1	18	12	36	16,316,704
	mit_32_d507	5,444	3	126	<1	40	17	187	8,965,766
	mit_76_studyroom	3,322	2	45	<1	47	19	114	15,837,152
	mit_dorm_next_sj	2,696	2	25	<1	27	7	62	3,665,537
	mit_lab_hj	1,906	1	15	<1	13	6	36	9,544,549

Table 1. Running times (in minutes) for all steps of the presented approach. Running times were measured on a workstation with an Intel Core i7-3770 3.5GHz CPU and 16GB of RAM.

along with complete infrastructure for rendering color and depth videos. Photorealistic color videos are produced with global illumination.

We augment the dataset in a number of ways. First, the original dataset released by Handa et al. provides four reference camera trajectories for each scene. However, these trajectories are short and do not model comprehensive scanning behaviors. The average reference trajectory is 39 seconds long and images only 45% of the surface area of the living room and 40% of the surface area of the office. Second, the provided noise model for the range camera is quite limited, yielding unrealistically clean depth images. One indication of the simplicity of the original trajectories and noise model is that pure visual odometry approaches perform very well, due to limited odometry drift and lack of complicated loop closures. The third limitation of the original release is the lack of a reference surface model for the office scene, which is represented procedurally; this precludes the evaluation of surface reconstruction accuracy on this scene.

We have adapted the dataset to evaluation of complete scene reconstruction pipelines. First, we have created two camera trajectories for each scene that model thorough handheld imaging for the purpose of comprehensive reconstruction. Table 2 lists the lengths and the surface area coverage rates of the trajectories. Second, we have integrated a comprehensive noise model that incorporates disparity-based quantization, realistic high-frequency noise, and a model of low-frequency distortion estimated on a real depth camera. This noise model has been previously used for surface reconstruction evaluation on synthetic data [12, 11]. Third, we have generated a dense point-based surface model for the office scene that enables the measurement of surface reconstruction accuracy. We have corresponded with the authors and verified that these extensions are in line with the intended usage of the dataset. The augmented dataset will be released upon publication.

Reconstructed models for two of the sequences are shown in Figures 3 and 4.

	Time (sec)	Length (meters)	Coverage (%)
Living room 1	96	37.2	94
Living room 2	78	36.8	85
Office 1	90	32.1	91
Office 2	85	38.1	83

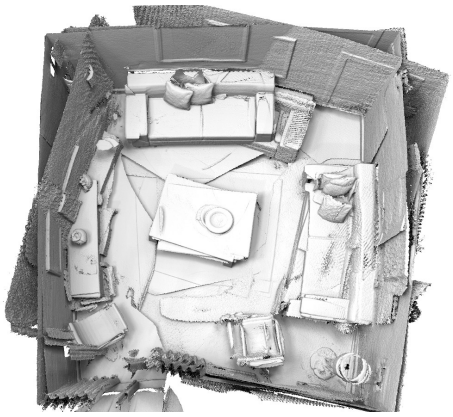
Table 2. Statistics for the augmented ICL-NUIM dataset. Camera trajectory duration, arc length, and surface coverage rates for the four simulated sequences.

Appendix E: Reconstruction accuracy

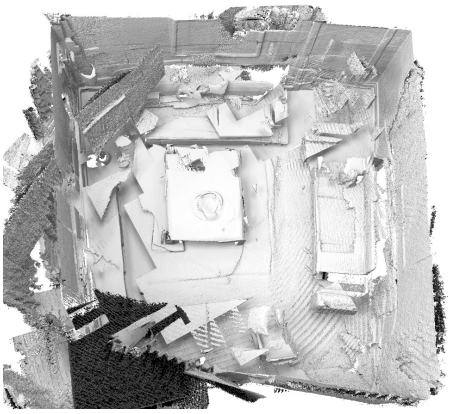
In this appendix we report additional measures of reconstruction accuracy on ICL-NUIM sequences. This supports Section 6.2 in the paper. Table 3 in the paper reports the mean distance of each reconstructed model to the ground-truth surface. Table 3 below reports the corresponding median distances. Our approach reduces the average median error by a factor of 2.7 relative to the closest alternative approach (SUN3D SfM). Table 4 reports the accuracy of the camera trajectories estimated by each pipeline, using the RMSE metric described by Handa et al. Our approach reduces the average error by a factor of 2.2 relative to the closest alternative approach.

	Kintinuous	DVO SLAM	SUN3D SfM	Ours	GT trajectory
Living room 1	0.17	0.16	0.08	**0.03**	0.03
Living room 2	0.10	**0.05**	0.06	**0.05**	0.02
Office 1	0.10	0.08	0.11	**0.02**	0.01
Office 2	0.09	0.07	0.06	**0.03**	0.02
Average	0.12	0.09	0.08	**0.03**	0.02

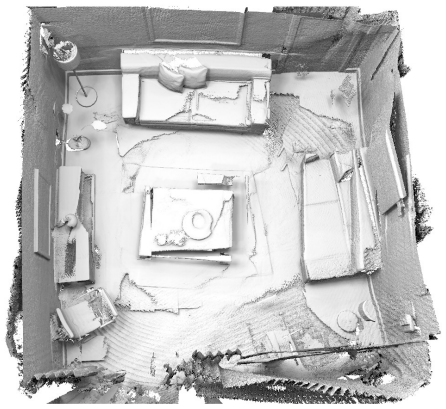
Table 3. Surface reconstruction accuracy on ICL-NUIM sequences. Median distance of each reconstructed model to the ground-truth surface, in meters.



(a) Kintinuous



(b) DVO SLAM



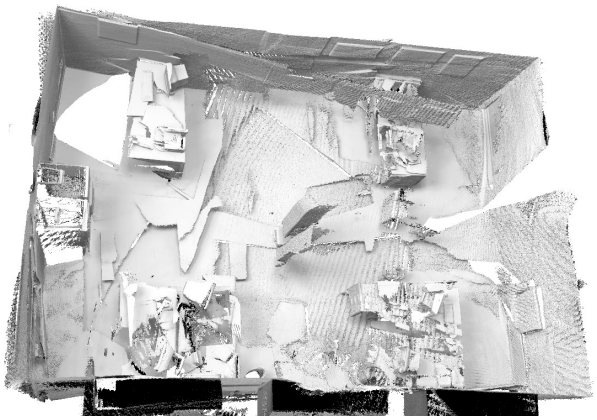
(c) SUN3D SfM

(d) Our approach

(e) Optional non-rigid refinement

Figure 3. Reconstruction of the Living room 1 sequence from the augmented ICL-NUIM dataset. (a) Reconstruction produced by Kintinuous [9]. (b) Reconstruction produced by DVO SLAM [4]. (c) Reconstruction produced by the RGB-D structure-from-motion pipeline of Xiao et al. [10]. (d) Reconstruction produced by our approach. (e) An optional non-rigid refinement of our result using SLAC [11].

(a) Kintinuous



(b) DVO SLAM

(c) SUN3D SfM

(d) Our approach

(e) Optional non-rigid refinement

Figure 4. Reconstruction of the Office 1 sequence from the augmented ICL-NUIM dataset. (a) Reconstruction produced by Kintinuous [9]. (b) Reconstruction produced by DVO SLAM [4]. (c) Reconstruction produced by the RGB-D structure-from-motion pipeline of Xiao et al. [10]. (d) Reconstruction produced by our approach. (e) An optional non-rigid refinement of our result using SLAC [11].

	Kintinuous	DVO SLAM	SUN3D SfM	Ours
Living room 1	0.27	1.02	0.21	**0.10**
Living room 2	0.28	0.14	0.23	**0.13**
Office 1	0.19	0.11	0.24	**0.06**
Office 2	0.26	0.11	0.12	**0.07**
Average	0.25	0.35	0.20	**0.09**

Table 4. Accuracy of estimated camera trajectories (RMSE).

Appendix F: Perceptual evaluation procedure

This appendix describes the experimental procedure used for quantitative evaluation of reconstruction quality on real-world scenes. This supports Section 6.3 in the paper. Our experimental design is based on pairwise comparisons, which are commonly used for quantitative evaluation of computer graphics techniques in the absence of ground-truth measurements [5, 3, 6, 1]. The pairwise comparison interface is demonstrated in the supplementary video.

The interface shows a short reference clip from an input color video. Below the reference video, two corresponding renderings of different reconstructions are shown side by side, in random left-right order. Each video shows a colored reconstruction of the scene rendered along the corresponding camera trajectory. The task is to indicate which of the two clips is more similar to the reference (or choose neither).

Each ICL-NUIM and SUN3D sequence was tightly covered with randomly sampled timestamps that are at least 10 seconds apart. For each sampled timestamp, a 1.5-second snippet of the original color video and a corresponding rendering of each colored model in the comparison set were produced.

Experiments were conducted using Amazon Mechanical Turk (MTurk). A single Human Intelligence Task (HIT) comprised all pairwise comparisons for a single sampled timestamp, along with four control questions that check for consistency and correctness. To check for consistency, two control questions replicate randomly chosen comparisons in the HIT with flipped left-right order. The other two control questions check for correctness by showing comparisons for which the correct answer is unambiguous. Each HIT is preceded by an introductory screen in which the task is explained. We rejected HITs for which the answers to two or more control questions were incorrect. Workers were paid 10¢ per HIT. All HITs were duplicated 20 times.

The experimental procedure was used to collect pairwise comparisons and compute BRE scores [8] for the ICL-NUIM sequences. We evaluated models reconstructed by Kintinuous, DVO SLAM, SUN3D SfM, our approach, and by integration of the noisy input data along the ground-truth trajectory. 32 timestamps were sampled from the four sequences. Each HIT comprised $\binom{5}{2}$ comparisons among the five reconstructed models, plus 4 control questions for a total of 14 comparisons per HIT. 640 HITs were deployed. A total of 71 unique workers performed 8,960 pairwise comparisons. 8% of the HITs were rejected based on the controls. Excluding control questions and "About the same" responses (see supplementary video), 5,083 pairwise comparisons yielded strict preferences that contributed to the computation of BRE scores.

The same experimental procedure was applied to the eight SUN3D sequences. We evaluated models reconstructed by Kintinuous, DVO SLAM, SUN3D SfM, and our approach, along with the manually-assisted reconstructions provided by Xiao et al. 63 timestamps were sampled from the eight sequences. Each HIT consisted of 14 pairwise comparisons including controls. 1,260 HITs were deployed and a total of 114 unique workers performed 17,640 comparisons. 14% of the HITs were rejected. Excluding control questions and "About the same" responses, 8,400 pairwise comparisons yielded strict preferences that contributed to the computation of BRE scores.

References

[1] E. Garces, A. Agarwala, D. Gutierrez, and A. Hertzmann. A similarity measure for illustration style. *ACM Transactions on Graphics*, 33(4), 2014. 7

[2] 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

[3] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A probabilistic model for component-based shape synthesis. *ACM Transactions on Graphics*, 31(4), 2012. 7

[4] C. Kerl, J. Sturm, and D. Cremers. Dense visual SLAM for RGB-D cameras. In *IROS*, 2013. 2, 3, 5, 6

[5] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics*, 29(4), 2010. 7

[6] P. O'Donovan, J. Libeks, A. Agarwala, and A. Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics*, 33(4), 2014. 7

[7] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, 2009. 1

[8] F. Wauthier, M. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *ICML*, 2013. 7

[9] 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, 3, 5, 6

[10] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, 2013. 2, 3, 5, 6

[11] Q.-Y. Zhou and V. Koltun. Simultaneous localization and calibration: Self-calibration of consumer depth cameras. In *CVPR*, 2014. 2, 3, 4, 5, 6

[12] Q.-Y. Zhou, S. Miller, and V. Koltun. Elastic fragments for dense scene reconstruction. In *ICCV*, 2013. 4