2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

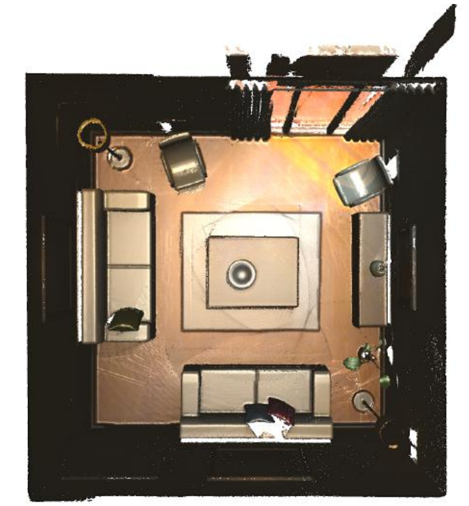
Surfel-Based Dense RGB-D Reconstruction with Global and Local Consistency

Yi Yang, Wei Dong, and Michael Kaess

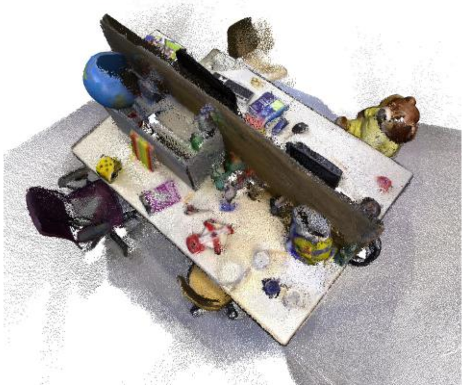
***Abstract*— Achieving high surface reconstruction accuracy in dense mapping has been a desirable target for both robotics and vision communities. In the robotics literature, simultaneous localization and mapping (SLAM) systems use RGB-D cameras to reconstruct a dense map of the environment. They leverage the depth input to provide accurate local pose estimation and a locally consistent model. However, drift in the pose tracking over time leads to misalignments and artifacts. On the other hand, offline computer vision methods, such as the pipeline that combines structure-from-motion (SfM) and multi-view stereo (MVS), estimate the camera poses by performing batch optimization. These methods achieve global consistency, but suffer from heavy computation loads. We propose a novel approach that integrates both methods to achieve locally and globally consistent reconstruction. First, we estimate poses of *keyframes* in the offline SfM pipeline to provide strong global constraints at relatively low cost. Afterwards, we compute odometry between *frames* driven by off-the-shelf SLAM systems with high local accuracy. We fuse the two pose estimations using factor graph optimization to generate accurate camera poses for dense reconstruction. Experiments on real-world and synthetic datasets demonstrate that our approach produces more accurate models comparing to existing dense SLAM systems, while achieving significant speedup with respect to state-of-the-art SfM-MVS pipelines.**

I. Introduction

Dense 3D reconstruction with accurate geometry is desired for different applications such as infrastructure inspection, indoor robot navigation, and virtual reality. The two key demands for a reconstruction algorithm to fulfill are: (1) accurate modelling of the global geometry of a large scale environment in terms of global geometric consistency, and (2) detailed scene description such as locally consistent shape, texture, and color. While many efforts in SfM and dense mapping in SLAM have been devoted to address both requirements [1]–[5], they often fail in one of the criteria.



(a) Augmented-ICL-lr1 [6] top view via splatted rendering



(b) TUM fr3 [8] point cloud office scene

Fig. 1. Reconstructions results for a synthetic and a real sequence

Specifically, in computer vision methods such as SfM, the initial structure and camera poses are computed using feature points. Later, bundle adjustment (BA) is applied to optimize the overall structure and poses by minimizing the reprojection error. One of the challenges in SfM is the presence of outliers in BA. In reality, the outliers from either the mismatched features or the incorrect pose initialization produce large error. BA distributes this error across all structures and poses, and results in inaccurate camera pose estimations. In addition, the reconstruction systems usually reject the image input in a featureless region, because feature-based pose estimation sometimes fails at these regions. On the other hand, systems that are dedicated to generating detailed local geometry and accurate short-term odometry in a room-size scene, such as ElasticFusion [5], often fail at large scale operations due to tracking failures. Other offline or online dense RGB-D reconstruction systems such as [6] and [7] produce fairly accurate models, but often at the cost of intensive computation and time.

In the paper, we propose a method that combines the advantages from SfM and SLAM to achieve high fidelity and accuracy in both local and global geometry in a large-scale 3D reconstruction using an RGB-D camera, as shown in Fig. I. This is achieved by using a factor graph-based

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {yiy4, weidong, kaess}@andrew.cmu.edu

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

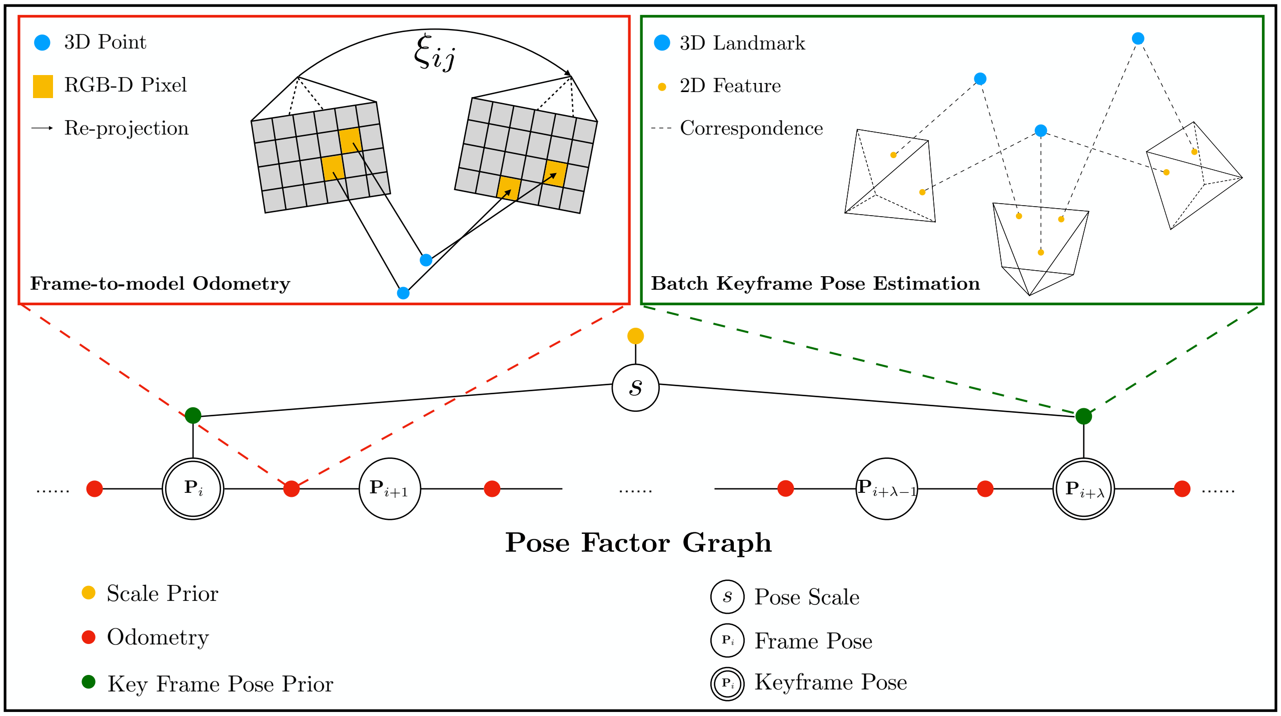


Fig. 2. Frame-to-model odometry is solved using joint ICP and photometric optimization, as shown in the picture in the top left. The keyframe pose priors are solved using SfM, as shown in the top right. The factor graph in the bottom demonstrates how to combine the keyframe priors with the odometry.

optimization [9] of the camera poses obtained from both SfM and a dense mapping SLAM system. Specifically, a few frames are selected from the color images as keyframes, and processed using SfM algorithms such as OpenMVG [1], [10], Theia [11] or COLMAP [12]. Then, keyframes poses are interpolated and combined with the RGB-D camera tracking from ElasticFusion [5], a state-of-the-art dense mapping system, to recover the overall trajectory via factor graph optimization. Finally, an online surfel-based fusion method used in ElasticFusion is applied to generate the final model based on the optimized camera poses. Our method uses the RGB-D odometry to enhance the local geometry consistency, and SfM to maximize the global geometry consistency.

There are two main contributions of this paper:

- We introduce the framework to combine SfM and SLAM in a factor graph formulation. Although our method requires offline pose optimization, it produces more accurate model geometry comparing to the online methods, and is much less time-consuming comparing to other offline reconstruction systems.
- We detail our methods in combining the result from SfM and SLAM. We conduct both real world experiments and tests on a synthetic dataset to provide a thorough analysis of the system runtime and performance. Our evaluation includes comparisons to both offline SfM-MVS pipeline and online dense mapping. The results show that our method outperforms both online dense mapping and offline 3D reconstructions in various metrics.

II. Related Work

The problem of both local and global geometric consistency has been addressed in many large scale offline SfM-MVS pipelines and online dense mapping systems. One of the most popular approaches is based on the idea of divide-and-conquer. Zhu et al. [13] and Yao et al. [14] have introduced the technique of locally readjusting the camera poses to improve the reconstruction quality. This method is effective in terms of smoothing out the rough surfaces and recovering more accurate local geometry. However, these methods still suffer from pre-defined heuristics such as the size of local camera group. Built upon these ideas, many state-of-the-art SfM pipelines such as COLMAP by Schönberger et al. [12] and OpenMVG by Moulon et al. [1] have developed complete systems that apply local refinement to recover better local reconstruction quality. These state-of-the-art systems provide globally accurate camera pose estimation. However, structure estimation from multi-view geometry is usually prone to outliers, and procedures such as patch matching in MVS are time-consuming. In addition, the resulting mesh from SfM-MVS pipeline usually requires post-processing that is not only time-consuming but also labor intensive.

As RGB-D sensors become popular, many offline systems are developed to use the depth information to generate highly accurate 3D model. For example, Zhou and Koltun [15] propose to construct small fragments of a large scene, and then use volumetric registration to combine these small fragments. Also, Zhou et al. [16] apply elastic regularization to merge all fragments. Apart from these methods, Choi et

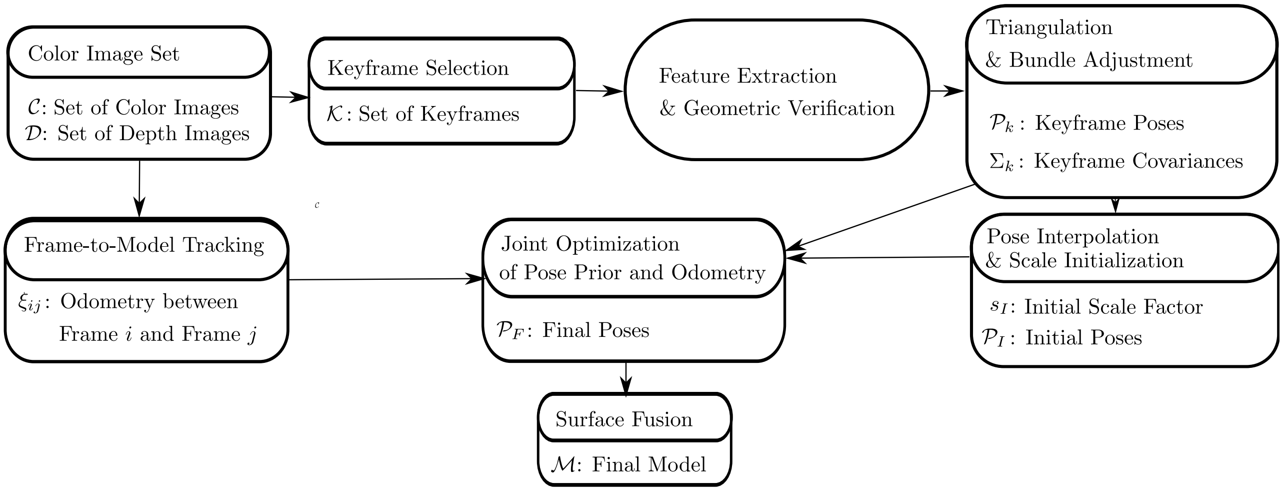

Color Image Set
$\mathcal{C}$: Set of Color Images
$\mathcal{D}$: Set of Depth Images
Keyframe Selection
$\mathcal{K}$: Set of Keyframes
Feature Extraction
& Geometric Verification
Triangulation
& Bundle Adjustment
$\mathcal{P}_k$: Keyframe Poses
Σ_k : Keyframe Covariances
c
Frame-to-Model Tracking
ξ_{ij} : Odometry between
Frame i and Frame j
Joint Optimization
of Pose Prior and Odometry
$\mathcal{P}_F$: Final Poses
Pose Interpolation
& Scale Initialization
s_I: Initial Scale Factor
$\mathcal{P}_I$: Initial Poses
Surface Fusion
$\mathcal{M}$: Final Model

Fig. 3. System block diagram. The top of each block shows the sub-components for each system, and the bottom of the block shows the output of the corresponding system. From a set of RGB-D images, a set of *keyframes* $\mathcal{K}$ is selected, and processed using the SfM pipeline. The resulting set of keyframe poses $\mathcal{P}_k$ and the corresponding covariance Σ_k are generated; these are treated as strong priors in the later optimization. Frame-to-model tracking provides odometry using joint optimization of ICP and photometric information. In preparation to the joint optimization of prior and odometry, keyframes are interpolated to provide a good initial value for the optimization. We optimize the poses incrementally as the new odometry becomes available, and fuse the new surfels from each frame into the existing model.

al. [6] apply a similar idea of constructing many overlapping small fragments, and then use Iterative Closest Point (ICP) to register the fragments together to complete a full scene. Recently, Open3D [17] introduces various algorithms to reconstruct the 3D environment with RGB-D information. However, these systems still face problems such as requiring manual alignment of the unregistered camera frames from separated fragments, and poor initialization of the camera poses could lead to incorrect convergence of the camera poses. These works produce accurate models, but at a cost of hours or even days of processing time. Different from these methods that use small fragment registration, our system focuses on providing accurate camera poses in dense reconstruction. It does not involve the computationally intensive registration.

Another category of the dense reconstruction is SLAM with RGB-D sensor. As a subject of great interest, systems such as KinectFusion [3], Kintinuous [4], InifiniTAM [18], [19] focus on the surface reconstruction of the model using Truncated Signed Distance Function (TSDF) [20] that stores the signed distance from every voxel to its nearest surface. Although these systems are able to achieve high performance in a room-size reconstruction, they rely on the depth information to conduct frame-to-model ICP, which is brittle when the depth measurements are subject to significant sensor noise. As a result, they are vulnerable to accumulated drift. ElasticFusion [5] incorporated the color information in the frame-to-model tracking to minimize the effect of ICP drift. However, the ICP drift still exists when the scene is large. Another approach applied by the state-of-the-art BundleFusion [7] is the SIFT feature correspondence search and BA that is similar to the works such as ORB-SLAM [21]. Although BundleFusion achieves superior performance in indoor reconstruction, it is computationally demanding, requiring two high-end GPUs to run in real-time. FlashFusion [22] applies a similar approach to use ORB [23] feature correspondence and keyframe integration without GPU. It achieves a similar performance to BundleFusion with much lower computation needs, but its keyframe integration strategy leads to a lower surface area coverage comparing to other dense mapping systems; it substantially sub-samples the frames and voxels. A common characteristic in all of the dense mapping methods in SLAM is that they often fail at a larger scene due to incorrect odometry estimation or incorrect loop closures. These failures often lead to misalignments in the final reconstruction. However, they provide highly accurate estimation of the change of poses between frames. We take advantage of the accurate short-term odometry estimation from the SLAM methods to achieve consistent local geometry.

III. Globally Consistent Pose Estimation using Factor Graph

Given a set of N RGB-D images captured by a commerical RGB-D camera, keyframe camera pose priors are obtained through an SfM pipeline. Later, the camera poses are jointly optimized by the prior and odometry from the dense tracking. The optimized camera poses are used in the final reconstruction. Fig. 3 shows an illustration of the system architecture.

A. Prior and Odometry Estimation

Keyframe Pose Estimation: The set of color images $\mathcal{C}$ is evenly down-sampled with an interval of λ. The resulting color maps are regarded as *keyframes* $\mathcal{K} \subset \mathcal{C}$. Using the state-of-the-art system COLMAP [12], we estimate the i-th frame pose $\mathbf{P}_i \in \mathrm{SE}(3)$ and its corresponding covariance $\Sigma_{\mathbf{P}_i}$. The resulting camera pose is regarded as the camera pose prior in the optimization in Sec. III-B. We refer to Schönberger et al. [2] and Moulon et al. [1] for a detail description of the SfM method in systems such as COLMAP and OpenMVG.

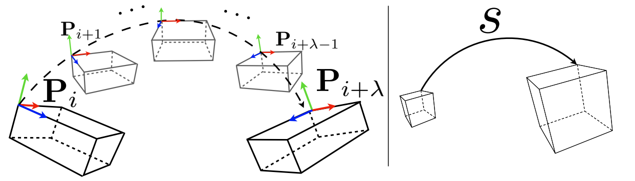


Fig. 4. (Left) Illustration of the pose interpolation on manifold. (Right) 3D similarity transformation.

Frame Odometry Estimation: The pose change between the i-th and j-th consecutive frames $\xi_{ij} \in \mathbb{R}^6$ is estimated using a joint frame-to-model [24] optimization. The cost function over the relative pose joins the point-to-plane distance and the photometric error, and is minimized in image pyramids for faster convergence. For details, we refer to the original ElasticFusion by Whelan et al. [5].

B. Factor Graph Representation

At the core of our proposed system is the joint optimization of the keyframe pose prior and frame-to-model odometry. To bring these together, we propose to use the factor graph formulation shown in Fig. 2. Specifically, given a keyframe $\mathbf{K}_i$ where $i \in \{1, 2, \cdots \lfloor \frac{N}{\lambda} \rfloor\}$ and its corresponding pose $\mathbf{P}_i$ estimated from SfM, we treat the keyframe pose as the pose prior in the factor graph. In addition, we represent the frame-to-model odometry ξ_{ij} between frame i and j as an odometry factor. Since the keyframe prior and the frame-to-model odometry are independently estimated, these two kinds of factors are not aligned in the same coordinate frame. In addition, due to the scale ambiguity issue in SfM, we need to simultaneously estimate the scale that best aligns the prior and the odometry. The connections of the prior factor and the odometry factor are shown in Fig. 2. Given this formulation, the optimal solution of the problem is the *maximum a posteriori* estimation of the factor graph:

$$\mathcal{X}^* = \arg\min_{\mathcal{X}} \sum_{i=1}^{N_K} \|r_s\|^2_{\sigma_s^2} + \sum_{i=1}^{N} \|\mathbf{r}_{\mathbf{P}_i}\|^2_{\Sigma_{\mathcal{P}_i}} + \sum_{i,j=1}^{N} \left\|\mathbf{r}_{\xi_{ij}}\right\|^2_{\Sigma_{\xi_{ij}}}, \tag{1}$$

where the set of optimal state is the set of all of the optimal poses and the optimal scale:

$$\mathcal{X}^* = \{\mathcal{P}^*, s^*\}. \tag{2}$$

r_s, $\mathbf{r}_{\mathbf{P}_i}$, and $\mathbf{r}_{\xi_{ij}}$ correspond to the residuals of scale, pose prior, and odometry respectively. σ_s^2, $\Sigma_{\mathbf{P}_i}$, and $\Sigma_{\xi_{ij}}$ correspond to the covariance of the scale, the pose prior, and the odometry. This problem can be solved using Levenberg-Marquardt non-linear optimization to iterate at the linearization point $\hat{\mathcal{X}}^k$. At the k-th iteration, the linearization can be expressed as the following Taylor expansion:

$$r_{s_i}(\hat{\mathcal{X}}^k + \delta\hat{\mathcal{X}}^{k+1}) \approx r_{s_i}(\hat{\mathcal{X}}^k) + H^k_{s_i}\delta\hat{\mathcal{X}}^{k+1}, \tag{3}$$

$$\mathbf{r}_{\xi_j}(\hat{\mathcal{X}}^k + \delta\hat{\mathcal{X}}^{k+1}) \approx \mathbf{r}_{\xi_j}(\hat{\mathcal{X}}^k) + H^k_{\xi_j}\delta\hat{\mathcal{X}}^{k+1}, \tag{4}$$

where we use the first order term to approximate the actual value. The measurement Jacobian H at the k-th iteration and the corresponding state update can be expressed as:

$$H^k_{s_i} = \left.\frac{\delta r_{s_i}}{\delta\mathcal{X}}\right|_{\mathcal{X}=\hat{\mathcal{X}}^k}, \; H^k_{\xi_{ij}} = \left.\frac{\delta \mathbf{r}_{\xi_{ij}}}{\delta\mathcal{X}}\right|_{\mathcal{X}=\hat{\mathcal{X}}^k}, \tag{5}$$

$$\hat{\mathcal{X}}^{k+1} = \hat{\mathcal{X}}^k \oplus \delta\mathcal{X}^{k+1}. \tag{6}$$

We use the $\oplus$ operator to represent retraction.

C. Pose Initialization: Interpolation on SE(3) Manifold

Initialization is of paramount importance to solve the non-convex optimization problem. It is important to initialize the state within close proximity of the optimal value to allow final convergence. If the sub-sampled interval λ is large, using the keyframe pose prior to initialize the intermediate poses is likely inaccurate, and thus might lead to wrong convergence. We propose to use an on-manifold interpolation based on the method shown in Žefran and Kumar [25] to initialize the poses in between two keyframes.

Let two keyframe poses be $\mathbf{P}_i$ and $\mathbf{P}_{i+\lambda}$ corresponding to the i-th and the $(i+\lambda)$-th frames, we want to find the intermediate pose $\mathbf{P}_k$ corresponding to the k-th frame that satisfies $i \leq k \leq i + \lambda$. This can be achieved by first finding the difference $\delta\mathbf{P}$ between two poses, and then map the difference from SE(3) to $\mathfrak{se}(3)$ represented by $\delta\xi$. Given the tangent space of the Lie manifold preserves linearity, we can estimate the intermediate poses corresponds to k as:

$$\mathbf{P}_k = \mathbf{P}_i \exp\left(\frac{k-i}{\lambda}\delta\xi\right). \tag{7}$$

D. Scale Initialization

Due to the scale ambiguity in SfM, we need to find the scale factor $s \in \mathbb{R}$ that best aligns the two sets of poses. Similar to the pose initialization in Sec. III-C, we also need to make sure that the scale is correctly initialized. We propose to use the 3D landmark $\mathbf{X}_{sfm}$ captured by the first frame in the SfM pipeline. Assuming that the first camera frame should be aligned with the global coordinate, we have the pose for the first frame expressed in the global coordinate frame to be $\mathbf{P} = \mathbf{I}_{4\times4}$. The set of first frame landmarks $\mathbf{X}_{sfm}$ corresponds to the set of 2D pixel coordinates $\mathbf{x}_1$. We can extract the corresponding 3D landmarks $\mathbf{X}_{slam}$ from the first frame depth image using the following camera projection function:

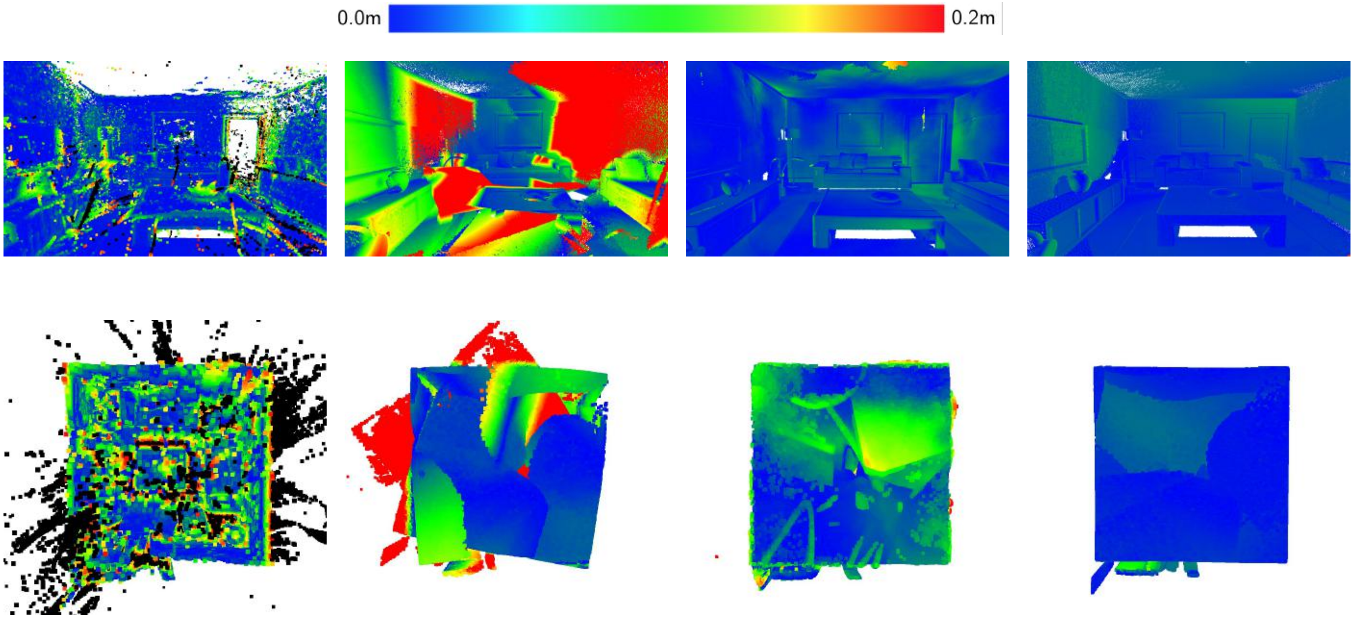
$$\mathbf{X}_{slam} = \pi^{-1}(\mathbf{P}_1 = \mathbf{I}_{4\times4}, \mathbf{x}_1), \tag{8}$$

where $\pi^{-1}(\cdot,\cdot)$ is the inverse camera projection that maps the pixel coordinate to the world coordinate. Then, based on the method by Horn [26], we estimate a similarity transformation between two sets of 3D points using singular value decomposition:

$$\mathbf{U}, \mathbf{S}, \mathbf{V} = \text{svd}(\tilde{\mathbf{X}}_{sfm} \otimes \tilde{\mathbf{X}}_{slam}), \tag{9}$$

$$\mathbf{R} = \mathbf{U}\mathbf{V}^\top, \tag{10}$$

$$\tilde{\mathbf{X}}_{aligned} = \mathbf{R}\tilde{\mathbf{X}}_{sfm}. \tag{11}$$



COLMAP-MVS ElasticFusion Redwood Proposed

Fig. 5. Heatmaps showing reconstruction error on A-ICL-lr1. The top row and the bottom row are the internal view and the top view of the room scene, respectively. Points more than 0.2m away from the ground truth are removed.

We express the initial scale as:

$$s_I = \frac{\sum_{X_i \in \tilde{\mathbf{X}}_{slam}, X_j \in \tilde{\mathbf{X}}_{aligned}} X_i \cdot X_j}{\sum_{X_j \in \tilde{\mathbf{X}}_{sfm}} \|X_j\|^2}, \tag{12}$$

where the $\tilde{\mathbf{X}}_{sfm}$ and $\tilde{\mathbf{X}}_{slam}$ are the zero-centered 3D points stacked in a column matrix, and $\otimes$ denotes the outer product between two matrices. Since the depth image is subject to noise, the initial scale s_I is not entirely accurate. We refine the scale in the optimization as shown in Fig. 2.

E. Optimization and Surfel-based Model Reconstruction

Given the keyframe pose prior and the odometry are initialized and aligned, we optimize the camera poses using the incoming odometry. When a new odometry measurement is obtained using the method in Sec. III-A, it is inserted into the factor graph and incrementally optimized using iSAM2 [27]. Similar to ElasticFusion, the model is generated by surface splatting. Different from the surface deformation approach in ElasticFusion, our method relies on the globally consistent pose priors to constrain the input odometry. Therefore, we do not apply the surface loop closure and the model deformation.

IV. Experimental Results

A. Implementation

We implement the system based on two state-of-the-art systems, COLMAP and ElasticFusion. COLMAP is used as the SfM pipeline that generates the pose prior, and the tracking component in ElasticFusion is used to generate odometry. In addition, we adopt the surfel representation and the surface fusion strategy in ElasticFusion. We implement the proposed factor graph, Levenberg-Marquardt optimizer, and iSAM2 [27] optimizer using the GTSAM [28] library. Experiments are conducted on a Ubuntu 16.04 desktop with an Intel i7-6700 CPU and a GeForce GTX 1070 GPU.

B. Synthetic Dataset

We test our system on the Augmented ICL-NUIM (A-ICL) dataset with synthetic noise created by Choi et al. [6]. The A-ICL sequence is based on the original ICL-NUIM dataset by Handa et al. [29]. Both ICL and A-ICL sequences are created in the same synthetic living room scene, but A-ICL contains a longer camera trajectory (max of 2870 frames vs. max of 1510 frames) and a larger coverage of the scene. We compare the surface reconstruction accuracy against the ground truth (GT) model using CloudCompare [30], and we also compare the Root Mean Square of Absolute Trajectory Error (ATE-RMSE) using the tools from TUM-RGBD benchmark [8]. To demonstrate the performance of our system, we compare against the systems from three different categories: (1) offline SfM-MVS systems represented by COLMAP-MVS [12], (2) offline RGB-D reconstruction systems represented by Redwood [6], and (3) online dense mapping systems represented by ElasticFusion [5] and InfiniTAMv3 [19]. The surface reconstruction accuracy, system runtime, and absolute trajectory error are used as metrics to evaluate the system performance. Table I shows the results of the A-ICL living room dataset with noise. A heatmap visualization of the surface reconstruction accuracy is shown in Fig. 5.

Table I shows the mean and the standard deviation of the distance to the GT model. Although our surface reconstruction accuracy is slightly lower than the models from COLMAP-MVS and the Redwood, the model produced by our method has the lowest standard deviation and trajectory error. A collection of heatmaps illustrating the surface

TABLE I

RESULTS ON THE SYNTHETIC A-ICL DATASET

	Mean distance to GT model (cm)		Std. Dev. distance to GT model (cm)		System Runtime (min)		ATE RMSE (cm)	
	A-ICL-lr1	A-ICL-lr2	A-ICL-lr1	A-ICL-lr2	A-ICL-lr1	A-ICL-lr2	A-ICL-lr1	A-ICL-lr2
COLMAP-MVS	**1.59**	**1.35**	4.58	4.20	212.66	159.41	5.68	18.78
Redwood	3.00	2.02	2.98	1.83	∼300	∼300	N/A	N/A
ElasticFusion	7.71	7.78	6.91	6.34	(online)	(online)	66.61	28.53
InfiniTAMv3	7.33	10.25	6.39	6.87	(online)	(online)	46.07	73.64
Proposed	1.74	2.26	**1.34**	**1.78**	**29.37**	**6.93**	**5.14**	**3.79**

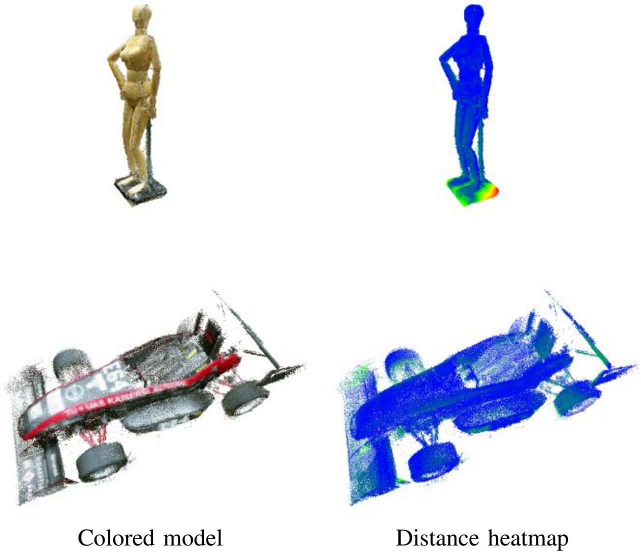


Fig. 6. Results on the real-world CoRBS dataset. Top row, *Human* sequence. Bottom row, *Racing Car* sequence. The output distance heatmap shows our method reconstructs accurate models.

reconstruction accuracy of A-ICL-lr1 is shown in Fig. 5. Furthermore, our method not only achieves a higher accuracy than the online methods, but also finishes the reconstruction in a much shorter period comparing to other offline methods.

C. Real World Dataset

We conduct additional tests on the CoRBS [31] real world dataset with surface ground truth model collected by a 3D scanner with sub-millimeter accuracy. This dataset is collected using a Kinect V2 hand-held RGB-D camera, and it contains a real racing car and a human-size model. The surface reconstruction accuracy of each object is shown in Table II, and the reconstructed models and the corresponding heatmaps are shown in Fig. IV-C.

TABLE II

RESULTS ON THE REAL-WORLD CORBS DATASET

	Mean distance (cm)		Std. dev. distance (cm)	
	Racing car	Human	Racing car	Human
ElasticFusion	6.07	52.98	4.45	49.0
Proposed	**1.30**	**1.12**	**1.15**	**1.84**

V. CONCLUSION

We present a dense reconstruction system that combines the advantages from both SfM and SLAM to recover both locally and globally consistent 3D models using an RGB-D sensor. We achieve performance by extracting keyframes and obtaining their corresponding pose priors from SfM, and locally adjusting the camera using odometry from dense SLAM. Our method shows better performance than the state-of-the-art dense SLAM system for both synthetic and real-world datasets, while providing much shorter processing time and comparable quality than other offline dense 3D reconstruction systems. In addition to the discussion about SfM and SLAM, the idea of combining strong camera priors and odometry can also be applied in situations such as GPS input as prior, or IMU input as odometry.

For future work, we would like to develop a solution to various parameter choices. For example, the number of keyframes should be automatically adjusted when the reconstruction data becomes larger, and more keyframes should be extracted when the camera motion is large. In addition, we are working towards an online system that is capable of achieving similar reconstruction accuracy.

ACKNOWLEDGMENT

We would like to thank Thomas Whelan for his help with ElasticFusion. We would also like to thank Ming Hsiao, Eric Westman, and Jerry Hsiung for valuable discussions.

REFERENCES

[1] P. Moulon, P. Monasse, and R. Marlet, “Adaptive structure from motion with a contrario model estimation,” in *Asian Conf. on Computer Vision (ACCV)*, Nov. 2012, pp. 257–270.

[2] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, “Pixelwise view selection for unstructured multi-view stereo,” in *Eur. Conf. on Computer Vision (ECCV)*, Oct. 2016, pp. 501–518.

[3] 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 *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, Oct. 2011, pp. 127–136.

[4] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald, “Kintinuous: Spatially extended KinectFusion,” *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Jul. 2012.

[5] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison, “ElasticFusion: Dense SLAM without a pose graph,” in *Robotics: Science and Systems (RSS)*, Rome, Italy, Jul. 2015.

[6] S. Choi, Q. Y. Zhou, and V. Koltun, “Robust reconstruction of indoor scenes,” in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Jun. 2015, pp. 5556–5565.

[7] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface re-integration," *ACM Transactions on Graphics*, vol. 36, no. 4, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3054739

[8] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2012, pp. 573–580.

[9] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, 2017. [Online]. Available: http://dx.doi.org/10.1561/2300000043

[10] P. Moulon and P. Monasse, "Unordered feature tracking made fast and easy," *ACM SIGGRAPH European Conference on Visual Media Production*, p. 1, Dec. 2012.

[11] C. Sweeney, T. Höllerer, and M. Turk, "Theia: A fast and scalable structure-from-motion library," in *Proceedings of the 23rd ACM Intl. Conf. on Multimedia*, Oct. 2015, pp. 693–696.

[12] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Jun. 2016, pp. 4104–4113.

[13] S. Zhu, T. Fang, J. Xiao, and L. Quan, "Local readjustment for high-resolution 3D reconstruction," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Jun. 2014, pp. 3938–3945.

[14] Y. Yao, S. Li, S. Zhu, H. Deng, T. Fang, and L. Quan, "Relative camera refinement for accurate dense reconstruction," in *Intl. Conf. on 3D Vision*, Oct. 2018, pp. 185–194.

[15] Q. Y. Zhou and V. Koltun, "Dense scene reconstruction with points of interest," *ACM Transactions on Graphics*, pp. 112:1–112:8, Jul. 2013.

[16] Q. Y. Zhou, S. Miller, and V. Koltun, "Elastic fragments for dense scene reconstruction," in *Intl. Conf. on Computer Vision (ICCV)*, Dec. 2013, pp. 473–480.

[17] Q. Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3d data processing," *Computing Research Repository*, vol. abs/1801.09847, 2018. [Online]. Available: http://arxiv.org/abs/1801.09847

[18] O. Kahler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. H. S. Torr, and D. W. Murray, "Very high frame rate volumetric integration of depth images on mobile device," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1241–1250, Nov. 2015.

[19] V. A. Prisacariu, O. Kähler, S. Golodetz, M. Sapienza, T. Cavallari, P. H. Torr, and D. W. Murray, "InfiniTAM v3: A framework for large-scale 3D reconstruction with loop closure," *ArXiv e-prints*, Aug. 2017.

[20] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *SIGGRAPH*, 1996, pp. 303–312. [Online]. Available: http://doi.acm.org/10.1145/237170.237269

[21] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardös, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[22] H. L and L. F, "FlashFusion: Real-time globally consistent dense 3D reconstruction using CPU computing," in *Robotics: Science and Systems (RSS)*, Jun. 2018.

[23] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Intl. Conf. on Computer Vision (ICCV)*, Nov. 2011, pp. 2564–2571.

[24] T. Whelan, H. Johannsson, M. Kaess, J. J. Leonard, and J. McDonald, "Robust real-time visual odometry for dense RGB-D mapping," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2013, pp. 5724–5731.

[25] M. Žefran and V. Kumar, "Interpolation schemes for rigid body motions," *Computer-Aided Design*, vol. 30, no. 3, pp. 179–189, Mar. 1998.

[26] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society America A*, vol. 4, pp. 629–642, 1987.

[27] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Intl. J. of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012.

[28] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," *Georgia Tech Technical Report*, 2012.

[29] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2014, pp. 1524–1531.

[30] "Cloudcompare user manual," https://www.danielgm.net/cc/doc/qCC/CloudCompare%20v2.6.1%20-%20User%20manual.pdf.

[31] O. Wasenmüller, M. Meyer, and D. Stricker, "CoRBS: Comprehensive RGB-D benchmark for SLAM using Kinect v2," in *Winter Conf. on Application of Computer Vision (WACV)*, Mar. 2016, pp. 1–7.