Noise-Resilient Reconstruction of Panoramas and 3D Scenes Using Robot-Mounted Unsynchronized Commodity RGB-D Cameras

SHENG YANG, BNRist, Tsinghua University, China
BEICHEN LI, Massachusetts Institute of Technology
YAN-PEI CAO, BNRist, Tsinghua University, China
HONGBO FU, City University of Hong Kong, Hong Kong
YU-KUN LAI, Cardiff University, UK
LEIF KOBBELT, RWTH Aachen University, Germany
SHI-MIN HU, BNRist, Tsinghua University, China

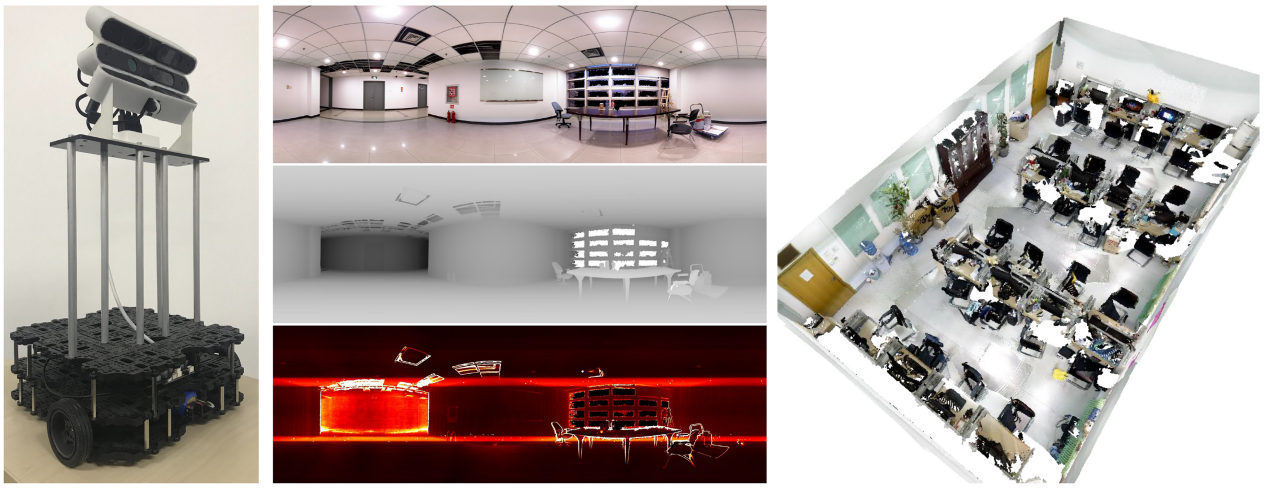


Fig. 1. We propose a two-stage approach for noise-resilient 3D reconstruction of large-scale indoor scenes through panoramic scanning. Left: We use a Turtlebot3 assembled with three unsynchronized commodity RGB-D sensors to perform multiple in-place rotations at different scanning positions. Middle: The first stage constructs 360° 3D panoramas (color, depth, and depth uncertainty) from unsynchronized RGB-D streams. Right: The second stage registers and stitches multiple panoramas into a globally consistent point cloud taking the pixel-wise uncertainty into account.

This work was supported by the National Key Technology R&D Program (Project Number 2017YFB1002604), the Joint NSFC-DFG Research Program (Project Number 61761136018), the Natural Science Foundation of China (Project Number 61521002), the Centre for Applied Computing and Interactive Media (ACIM) of School of Creative Media, City University of Hong Kong, and the Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.
Authors' addresses: S. Yang, Y.-P. Cao, and S.-M. Hu (corresponding author), BNRist, Tsinghua University, Room 3-507, 3-523, Information Technology Building (FIT), Beijing, China, 100084; emails: {shengyang93fs, caoyanpei}@gmail.com, shimin@tsinghua.edu.cn; B. Li, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, The Stata Center, Building 32-312, 32 Vassar Street, Cambridge, MA 02139; email: beichen@mit.edu; H. Fu, School of Creative Media, City University of Hong Kong, Level 7, Run Run Shaw Creative Media Centre, 18 Tat Hong Avenue, Kowloon Tong, Hong Kong; email: hongbofu@cityu.edu.hk; Y.-K. Lai, School of Computer Science and Informatics, Cardiff University, S/3.06 Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA; email: Yukun.Lai@cs.cardiff.ac.uk; L. Kobbelt, Visual Computing Institute, RWTH Aachen University, Room 117, RWTH Aachen, Lehrstuhl für Informatik 8, Ahornstraße 55, 52074 Aachen; email: kobbelt@cs.rwth-aachen.de.

0730-0301/2020/06-ART152 $15.00
https://doi.org/10.1145/3389412

We present a two-stage approach to first constructing 3D panoramas and then stitching them for noise-resilient reconstruction of large-scale indoor scenes. Our approach requires multiple unsynchronized RGB-D cameras, mounted on a robot platform, which can perform in-place rotations at different locations in a scene. Such cameras rotate on a common (but unknown) axis, which provides a novel perspective for coping with unsynchronized cameras, without requiring sufficient overlap of their Field-of-View (FoV). Based on this key observation, we propose novel algorithms to track these cameras simultaneously. Furthermore, during the integration of raw frames onto an equirectangular panorama, we derive uncertainty estimates from multiple measurements assigned to the same pixels. This enables us to appropriately model the sensing noise and consider its influence, so as to achieve better noise resilience, and improve the geometric quality of each panorama and the accuracy of global inter-panorama registration. We evaluate and demonstrate the performance of our proposed method for enhancing the geometric quality of scene reconstruction from both real-world and synthetic scans.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; **Vision for robotics**; *Point-based models*;

Additional Key Words and Phrases: Panorama, reconstruction, SLAM, robotics

ACM Reference format:
Sheng Yang, Beichen Li, Yan-Pei Cao, Hongbo Fu, Yu-Kun Lai, Leif Kobbelt, and Shi-Min Hu. 2020. Noise-Resilient Reconstruction of Panoramas and 3D Scenes Using Robot-Mounted Unsynchronized Commodity RGB-D Cameras. *ACM Trans. Graph.* 39, 5, Article 152 (June 2020), 15 pages.
https://doi.org/10.1145/3389412

1 INTRODUCTION

Modern scene understanding and navigation tasks [Qi et al. 2017; Zhang et al. 2015] require databases of high-fidelity 3D scenes [Armeni et al. 2016; Dai et al. 2017a; Hua et al. 2016], which are mostly acquired through either *hand-held scanning* [Dai et al. 2017a; Hua et al. 2016] or *panoramic scanning* [Armeni et al. 2016; Ikehata et al. 2015; Mattausch et al. 2014]. *Hand-held scanning* techniques take Color and Depth (RGB-D) video streams as input and utilize modern dense reconstruction systems [Dai et al. 2017b; Newcombe et al. 2011] or visual-Simultaneous Localization and Mapping (SLAM) algorithms [Mur-Artal and Tardós 2017] for tracking and integrating sequential frames. *Panoramic scanning*, on the other hand, schedules the scanning process into multiple in-place rotations to construct 3D panoramas for progressive integration at different viewpoints [Wijmans and Furukawa 2017]. Compared with *hand-held scanning*, which requires continuous focus on regions with sufficient geometric or photometric features for robust tracking, *panoramic scanning*, where in-place rotations are easier to be tracked [Chang et al. 2017; Taylor et al. 2015], becomes a practical alternative for industrial or commercial applications [Armeni et al. 2016; Ikehata et al. 2015], even for upcoming automated scanning scenarios with the aid of a progressive discrete motion planning module.

A variety of techniques have been developed to construct 360° panoramas using such a panoramic scanning scheme, and based on their input and output image types (i.e., whether containing depth information or not), we categorize them into three classes, namely 2D-to-2D, 2D-to-3D, and 3D-to-3D. Although it is possible to use 2D RGB cameras to recover coarse depth information for canonical stitching and Virtual Reality / Augment Reality (VR/AR) applications [Hedman et al. 2017; Hedman and Kopf 2018], the resulting depth quality is usually not sufficient for high-fidelity 3D reconstruction. Current 3D-to-3D techniques based on a single RGB-D camera [Taylor et al. 2015] have limited Field-of-View (FoV) when the Degree-of-Freedom (DoF) of sensor motion is restricted (e.g., with 1-DoF rotation) and, hence, cannot cover most of entire spherical panoramas. This narrow FoV problem can be addressed by utilizing multiple RGB-D cameras (e.g., arranged vertically for horizontal rotation), which, however, introduces new issues with camera calibration and synchronization.

The first issue of panoramic scanning with multiple cameras is how to recover the relative poses of these RGB-D frames. A convenient method is to use external positioning sensors, as used by Matterport [Chang et al. 2017], to directly measure their poses. However, under the circumstances when no external positioning sensors are available (or it is hard to perform precise calibration for customized assembly), an alternative choice to technically solve this problem is to rely on visual features for tracking. If shutter synchronization is accessible, this issue can be relegated to a general visual SLAM or reconstruction problem [Dai et al. 2017b; Mur-Artal and Tardós 2017] by pre-stitching synchronized frames with camera extrinsics. Unfortunately, most commodity depth sensors (including Kinect and PrimeSense) do not support shutter synchronization, and forcibly grouping them by timestamps will cause misalignments (Figure 6) due to neglected motions during shutter intervals. In addition, although some approaches for unsynchronized RGB cameras have proposed to utilize overlapped scanned areas as mutual information [Cadena et al. 2016], following their strategies to enlarge these areas to track multiple RGB-D cameras would easily cause severe depth interference and reduced FoV. Therefore, when neither external auxiliary hardware nor sufficiently overlapped areas are available, we need a new strategy for jointly solving poses. Otherwise, the resulting frames toward featureless areas (e.g., ceiling and ground) would eventually cause tracking loss [Yang et al. 2019].

The second issue is the inherent sensor noise, which is not severe when using high-quality laser scanners such as Faro 3D [Ikehata et al. 2015; Wijmans and Furukawa 2017] but becomes critical on commodity RGB-D frames [Cao et al. 2018]. Previous works handle noise during frame integration for continuous scanning through several general data structures, such as the Truncated Signed Distance Function (TSDF) volume [Dai et al. 2017b; Newcombe et al. 2011], the Probabilistic Signed Distance Function (PSDF) volume [Dong et al. 2018], and Surfels [Keller et al. 2013; Weise et al. 2009; Whelan et al. 2015b]. But only a few of them have further considered the influence of noise during frame registration [Cao et al. 2018; Dong et al. 2018]. Also, using the above data structures for constructing panoramas is both memory inefficient and computationally expensive. In addition, modeling the noise of depth measurements after panorama construction is important, since the subsequent steps for the inter-panorama registration and final integration are all affected by such uncertain measurements. Hence, how to represent scanned data and model their noise through an efficient and suitably organized structure is also an important task during the panorama construction process.

To address these issues and thus reconstruct high-fidelity 360° panoramas and 3D scenes represented by a point cloud, we present a novel approach suitable for tracking *unsynchronized* RGB-D cameras during panorama construction, with the noise of depth measurements modeled and further handled.

For the first issue, our strategy to achieve collaborative scanning is based on the consensus of motion of these cameras driven by an in-place rotator (e.g., a robot). Considering the coaxiality of their motion enables us to jointly derive their states without depending on synchronization or significant landmark co-occurrences. This is achieved through several novel regularization constraints under a factor graph optimization framework [Grisetti et al. 2010].

For the second issue, we choose the equirectangular image format for fusing color and depth measurements rather than those general data structures, so as to efficiently organize and estimate per-pixel uncertainty in the panorama domain (Figure 1, middle). With such an organized image structure and its noise models, we optimize the geometric quality of the reconstructed panorama regarding the data consistency in such an image domain, and further consider the influence of noise quantitatively during the subsequent inter-panorama registration and final integration (Figure 1, right).

In summary, our work makes three contributions. Firstly, we develop a feasible workflow that progressively reconstructs 3D panoramas and scenes, achieving higher accuracy than state-of-the-art reconstruction algorithms. Secondly, we propose a solution for jointly tracking unsynchronized cameras by formulating their motion consistency, without relying on significant visual co-occurrences and shutter synchronization. Thirdly, our approach infers pixel-wise depth measurement uncertainties in the equirectangular image domain, and further considers these uncertainties during subsequent operations, to enhance the geometric quality of panoramas and the final scene.

We assembled our prototype system (Figure 1, left) inspired by the pioneering *Matterport* [Chang et al. 2017; Matterport Inc. 2019] system. The only preparation for performing the proposed reconstruction is an initial calibration between cameras, while the relative transformations between frames and the rotation axis are automatically computed during scanning. To evaluate the effectiveness of the proposed approach quantitatively, we additionally perform experiments on several simulated scans from synthetic scenes.

2 RELATED WORK

We next briefly review approaches designed for the two stages of panoramic scanning: panorama construction (Section 2.1) and panorama integration (Section 2.2).

2.1 Panorama Construction

The key problem of image stitching for 2D-to-2D panorama construction, i.e., aligning and integrating multiple RGB frames, has been well-studied in the vision communities [Szeliski 2006]. Homography [Zaragoza et al. 2013] and deghosting approaches [Zhu et al. 2018] are two common and complementary solutions, with the same goal to reduce artifacts. Such 2D-to-2D methods can be directly extended to consider depth measurements as an additional channel, but they would cause misalignment when the parallax exists, since the used homography is essentially for mapping the same planar surface between images.

Hence, recent 2D-to-3D algorithms tend to predict depth information from input images and utilize a 6-DoF relative transformation rather than the homography for stitching. On condition of sufficient visual correspondences, recent approaches [Klingner et al. 2013; Schönberger et al. 2016] use Structure-from-Motion (SfM) [Snavely et al. 2006] for predicting relative camera poses. Based on these estimated poses, Multi-View Stereo (MVS) approaches such as plane sweeping [Häne et al. 2014] and their variants [Hedman et al. 2017; Hedman and Kopf 2018; Im et al. 2016] are performed to densify depth predictions on the images. Among them, Hedman et al. [2017] augment the near envelope for discouraging nearby depth hypotheses and achieve state-of-the-art depth quality sufficient for Virtual Reality / Augment Reality (VR/AR) panoramic applications, but such predicted depth information is still not precise enough for our purpose.

Aiming at high-fidelity dense reconstruction, various algorithms proposed for RGB-D frames mainly concentrate on precisely tracking sensors and refining depth measurements [Choi et al. 2015; Dai et al. 2017b; Whelan et al. 2015b; Zhou et al. 2013]. Specifically for 3D-to-3D panoramic scanning, Taylor et al. [2015] attempted to perform panoramic reconstruction through a single RGB-D sensor. Compared to their setup and approach, our scheme has two important advantages: (1) Our algorithm takes the input from multiple cameras and jointly optimizes their trajectories to achieve globally consistent stitching. (2) The reprojection parameters for stitching frames onto the reconstructed panorama are simultaneously solved with the poses of involved frames to predict the exact location of the rotation axis, while their method relies on the Manhattan assumption to use axis-aligned lines and surfaces in a scene for addressing the gravity orientation after these poses are determined. However, in practice, the gravity orientation does not precisely coincide with the direction of the rotation axis (see also Section 6.2 for a quantitative comparison).

Our scanning style remains the same as *Matterport* [Chang et al. 2017; Matterport Inc. 2019], a commercial system that uses external sensors for localizing their scanned frames. But technically, our system is extricated from the reliance on auxiliary devices by implementing visual-based localization for unsynchronized cameras, i.e., achieving the same goal with fewer hardware requirements. Furthermore, as demonstrated in our experiments (Sections 6.2–6.3), the constructed pixel-wise depth uncertainty models as an augmented channel of the panorama can enhance the quality of the panorama construction.

2.2 Aligning and Integrating Panoramas

Due to the discretization of scanning positions in such a panoramic scheme, constructed panoramas need to be jointly aligned for compositing final scenes. In geometric processing, this is referred to as model registration and typically accomplished through a coarse-to-fine procedure. In the coarse stage, sparse transformation between models can be acquired manually by user hints [Ikehata et al. 2015; Mura et al. 2014], or automatically through 2D/3D keypoint matching [Chang et al. 2017]. In the fine stage, direct methods based on photometric [Whelan et al. 2015a] or geometric [Besl and McKay 1992; Ren et al. 2019; Segal et al. 2009] costs are proposed to establish and optimize dense correspondences between two models. Typically for integrating 360° 3D panoramas, Taylor et al. [2015] propose to use projective association through equirectangular projection for efficient correspondence searching, and Wijmans et al. [2017] utilize an additional floor plan image to accomplish globally consistent alignment for high-quality laser scans. However, for low-cost depth sensors, dealing with sensor noise is critical or even necessary for obtaining high-quality reconstruction. From this perspective, our method also differs from Taylor et al. [2015] in the use of our obtained noise models, which support noise-aware inter-panorama registration as well as the final integration, thus effectively improving the accuracy of reconstructed scenes.

3 ASSUMPTIONS AND OVERVIEW

Our method is based on a scanning platform carrying multiple unsynchronized RGB-D cameras. Before scanning, we mount an additional fish-eye camera to perform a joint calibration [Maye et al. 2013] on all color and depth sensors to obtain their intrinsic and extrinsic parameters. Then we use Calibrating Localizing

and Mapping Simultaneously (CLAMS) [Teichman et al. 2013] to undistort raw depth measurements. During scanning, the platform is required to perform in-place rotations at multiple viewpoints. Specifically for wheeled robots, this can be implemented by setting the same speed in different directions for its two drive wheels. For each RGB-D sensor, we use the calibrated extrinsic parameters to map its depth frames to their corresponding color frames; thus, the input to our algorithm is a set of timestamped RGB-D images clustered by different scanning positions.

Our reconstruction algorithm is based on the following two pre-requisites: (1) The rotation of all cameras should be performed *coaxially* with a static axis, which requires a flat ground according to our assembly (rugged scenes can be adapted with an additional pan for performing stable rotation). (2) The rotation should also be performed *smoothly*, which means the angular acceleration should remain low during scanning. Although the coaxiality and smoothness pre-requisites may not be perfectly satisfied due to the unpredictable robot shaking, our method is tolerant to these practical phenomena with configurable parameters.

The reconstruction process is performed in two stages, namely *panorama construction* (Section 4), which stitches in-place RGB-D streams to individual 360° panoramas containing color, depth, and an additional per pixel depth uncertainty (Figure 1, middle), and *panorama integration* (Section 5), which registers and integrates these panoramas to form a complete 3D representation of the scene (Figure 1, right).

Each stage contains two phases: namely, pose estimation and data fusion. In the pose estimation phase (Section 4.1) of the first stage, our algorithm performs bundle adjustment with additional regularizations considering motion consistency to solve for the poses of frames and the rotation axis. Then, in its data fusion phase (Section 4.2), we warp raw color and depth measurements into an equirectangular representation of a desired panorama for statistical modeling of sensor noise, and perform an optimization to refine its geometric quality. In the second stage, the pose estimation phase is accomplished by consistently aligning spatially related panoramas through geometric correspondences to account for their mixed uncertainty, combining both the original surface distribution uncertainty and our extracted measurement uncertainty (Section 5.1). Finally, in its data fusion phase, based on such mixed uncertainty knowledge, the estimated poses of these aligned panoramas, as well as their uncertain depth measurements, are used to revise the final dense point cloud as a representation for the whole scene (Section 5.2).

4 PANORAMA CONSTRUCTION

4.1 Consistency Regularization for Panoramic Scanning

We address the challenge of jointly solving for the poses of frames and the rotation axis, by utilizing the characteristics of these coaxial rotations. Since all cameras and the axis constitute a fixed body and move together during scanning, we can thus use a unified physical model and extrinsics to describe their motion. Especially for such an in-place rotation, once these extrinsics between the axis and multiple cameras are solved, the status of these cameras can be parameterized through only 1-DoF, as the azimuth angle of the rotator. As shown in Figure 2 (left), the edges between the rotation axis and these cameras enable the regularization of

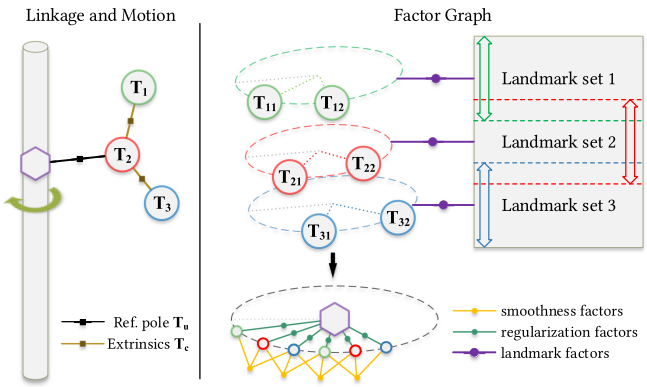


Fig. 2. Schematic of linkages and factors involved in our pose estimation problem. Left: The rotation (whose axis is illustrated as a cylinder) driven by the rotator is conducted through their static links. The left link in black stands for the transformation from the axis to the reference camera ($\mathbf{T}_u$), and the other links in brown represent the extrinsics ($\mathbf{T}_c$) for each RGB-D camera c. Right: Our underlying factor graph consists of three types of factors linking different variables. Bottom-Right: Through the conduction of sensor motions, additional factors can be established regarding their reference azimuth.

camera motions (as illustrated in the bottom-right circle in gray), where the azimuth is regarded as their point of reference. This enables us to solve for motions of unsynchronized cameras jointly without the requirement of sufficient landmark co-observations [Schmuck and Chli 2017].

In order to formulate this feature, we choose to use a factor graph framework [Grisetti et al. 2010] due to its flexibility of tackling multivariate optimization with various types of constraints. The underlying factor graph, denoted as $G = (\mathcal{X}, \mathcal{F}, \mathcal{E})$, consists of variable nodes $\mathcal{X}$ to be solved under the constraints of factors $\mathcal{F}$ via correspondences $\mathcal{E}$. We introduce four categories of variables, including:

(1) The classical landmark $y_j \in \mathcal{X}$ as defined in various visual SLAM approaches (e.g., Mur-Artal and Tardós [2017]), where $y_j \in \mathbb{R}^3$ represents the global position of landmark j;
(2) The augmented pose representation for frame i from camera c: $x_{ci} = \{\mathbf{T}_{ci}, \alpha_{ci}\} \in \mathcal{X}$, containing a traditional 6-DoF pose representation $\mathbf{T}_{ci} \in \mathbb{SE}^3$ and the proposed azimuth as $\alpha_{ci} \in [0, 2\pi)$, where $\mathbf{T}_{ci}$ denotes the pose of the frame i from the camera c w.r.t. the reference frame (without loss of generality, we choose the first received frame in our system as the reference frame, and its pose remains fixed during the optimization);
(3) The installation bias between the sensors and the rotation axis $\mathbf{T}_u \in \mathcal{X}$, where $\mathbf{T}_u \in \mathbb{SE}^3$ is denoted as the pose of the rotation pole w.r.t. the reference frame;
(4) The extrinsics between other cameras and the reference camera (the one that outputs the reference frame) $\mathbf{T}_c \in \mathcal{X}$, where $\mathbf{T}_c \in \mathbb{SE}^3$ stands for the 6-DoF pose of camera c in the coordinate system of the reference camera.

For the above rigid transformations, we use Euler angles to represent 3D rotations to facilitate the configuration of parameters w.r.t. imperfect robot motions. Compared with conventional factor

graph formulations for visual SLAM tasks [Mur-Artal and Tardós 2017], our proposed graph structure contains additional variables such as camera extrinsics (Category 4) and azimuth variables (Category 2) for online calibration and regularization, respectively. In fact, the variable for the pose of the rotation axis $\mathbf{T}_u$ constitutes the model-view transformation during the panorama construction phase (Section 4.2). Acquiring a correct transform for reprojecting measurements onto the panorama enables us to balance the contribution of each raw frame, i.e., they can generate consistent regions on the panorama image for calculating statistics.

Based on these variables, three types of factors are established (see Figure 2, right) as: (1) traditional landmark observation factors establishing the relations between frame poses and landmarks for bundle adjustment; (2) pose regularization factors regularizing camera motions to conform to horizontal rotations; (3) smoothness factors constraining consistent angular velocity between consecutive frames.

Landmark observation factors. Like previous works [Mur-Artal and Tardós 2017], our factor graph utilizes keypoint correspondences as our baseline of bundle adjustment, where the observation factor $\mathbf{f}_{ci,j}^{ob,\mathbb{V}} \in \mathcal{F}^{ob}$ for its corresponding frame x_{ci} and landmark y_j is defined as:

$$\mathbf{f}_{ci,j}^{ob,\mathbb{V}} \propto \exp\left(-\frac{1}{2}\|p_j^{\mathbb{V}} - \mathbf{K}_c(\mathbf{T}_{ci}^{-1}\cdot y_j)\|^2_{\Omega_{ci}^{\mathbb{V}}}\right), \tag{1}$$

where $\|\mathbf{e}\|^2_{\Omega} \triangleq \mathbf{e}^\top\Omega^{-1}\mathbf{e}$ is the squared Mahalanobis distance with the covariance matrix Ω. $p_j^{\mathbb{V}} = d_j[u_j, v_j, 1]^\top$ is the observation of y_j in the image coordinate $\mathbb{V}$ of x_{ci}, and $\mathbf{K}_c(\cdot)$ is the perspective projection function w.r.t. the intrinsic parameters of camera c.

We use the noise-aware bundle adjustment proposed by Cao et al. [2018] to deal with the uncertainty of raw depth measurements, where $\Omega_{ci}^{\mathbb{V}} = \mathrm{diag}(\sigma_u^2, \sigma_v^2, \sigma_d^2)$ is the covariance matrix capturing the confidence of independent measurements, with (σ_u^2, σ_v^2) given through the uncertainty from a keypoint extraction approach [Mur-Artal and Tardós 2017], and σ_d^2 assigned via the estimated variance from the depth model proposed by Handa et al. [2014].

Correspondences between frames and landmarks are built by extracting and comparing Oriented FAST and Rotation BRIEF (ORB) features [Mur-Artal and Tardós 2017]. We utilize RANSAC [Fischler and Bolles 1981] to sample and reach a consensus on confident 3D transformations for rejecting erroneous visual correspondences between frames from the same camera. Temporally distant frames are also examined for detecting and addressing loop closures through Randomized Ferns [Glocker et al. 2015].

In addition, correspondence search is carried out between different cameras through the temporally closest frames if there exist overlapping regions between these frames. As an example in our configuration, these overlapping regions are rather narrow (less than 2% of the whole image domain to alleviate depth interference), but sufficient for refining camera extrinsics online. For example, in our assembly, hundreds of frames and landmarks are used for solving the extrinsics $\mathbf{T}_c$ for two cameras.

Pose regularization factors. To make use of the consistency of motion and estimate the pose of the rotation axis, we introduce a regularization factor $\mathbf{f}_{ci}^{reg} \in \mathcal{F}^{reg}$ for each frame x_{ci} as follows:

$$\mathbf{f}_{ci}^{reg} \propto \exp\left(-\frac{1}{2}\|\mathbf{T}_{ci}^{-1}\cdot\mathbf{T}_u\cdot\mathbf{R}(\alpha_{ci})\cdot\mathbf{T}_u^{-1}\cdot\mathbf{T}_c\|^2_{\Omega^{reg}}\right), \tag{2}$$

where $\mathbf{R}(\alpha_{ci}) \in \mathbb{SO}^3$ is a pure azimuth rotation generated through α_{ci} for representing the state of rotation, and $\Delta\mathbf{T} \triangleq \mathbf{T}_u\cdot\mathbf{R}(\alpha_{ci})\cdot\mathbf{T}_u^{-1}$ reflects the expected pose state of the reference camera w.r.t. the reference frame according to the azimuth α_{ci}. Hence, the difference between the expectation of frame x_{ci} (which can be described as $\Delta\mathbf{T}\cdot\mathbf{T}_c$) and the estimation $\mathbf{T}_{ci}$ describes the severity of systematic shaking, and we use $\Omega^{reg} \in \mathbb{R}^{6\times6}$ as a configurable parameter for uniformly describing and considering such severity for all frames, whose translation and rotation parts are set according to the possible level of vibration determined by the hardware setup (see Section 6.1 for details). Such a redundancy for describing the expected and the actual poses, i.e., between variables α_{ci} and $\mathbf{T}_{ci}$, makes our algorithm tolerate imperfect rotations. When estimating the cost of such a factor, we linearize the overall transformation into a six-dimensional vector [Kümmerle et al. 2011].

Velocity smoothness factors. To promote uniformity of angular velocity, we establish smoothness factors $\mathbf{f}_i^{vel} \in \mathcal{F}^{vel}$ between adjacent frames as follows:

$$\mathbf{f}_i^{vel} \propto \exp\left(-\frac{1}{2}\|v_{i,i+1} - v_{i-1,i}\|^2_{\Omega^{vel}}\right), \tag{3}$$

where $v_{i,j} = (\alpha_i \ominus \alpha_j)/(t_i - t_j)$ is the angular velocity between two consecutive frames based on their azimuths α_i, α_j and timestamps t_i, t_j, with $\ominus$ denoting the wrap around subtraction with a modulo of 2π. $\Omega^{vel} \in \mathbb{R}$ again defines the confidence of such factors, whose values are assigned according to the stability of motion control (see Section 6.1). Since we define the angular velocity $v_{i,j}$ regardless of which camera it belongs to, we effectively avoid the requirement of hardware synchronization.

Optimization with robust kernels. Although landmark observations are filtered before being added into the factor graph, there may still exist erroneous correspondences. Hence, we additionally apply the Huber robust kernels [Latif et al. 2013] to all landmark observation factors, and define the overall optimization problem as:

$$\min_{X} \sum_{\mathcal{F}^{vel}} \mathbf{E}(\mathbf{f}_i^{vel}) + \sum_{\mathcal{F}^{reg}} \mathbf{E}(\mathbf{f}_{ci}^{reg}) + \sum_{\mathcal{F}^{ob}} \mathbf{H}\Big(\mathbf{E}(\mathbf{f}_{ci,j}^{ob,\mathbb{V}})\Big), \tag{4}$$

where $\mathbf{E}(\cdot) = -\log(\cdot)$ obtains the negative log-likelihood of these factor constraints in Equations (1)–(3), making their scale factors become irrelevant constants. $\mathbf{H}(\cdot)$ is the Huber cost function [Latif et al. 2013] for diminishing influences on incompatible pose observations.

In general, as discussed previously, our proposed structure of factor graph can be applied to similar panoramic scanning devices containing one or multiple cameras with smooth rotations. Detailed hardware and parameter configurations, and extensive experiments are given in Sections 6.1 and 6.2, respectively.

4.2 Constructing Panoramas and Noise Models

Processing in the panorama domain instead of using general data structures [Keller et al. 2013; Newcombe et al. 2011] can produce an organized image rather than a point cloud or a mesh, which is more

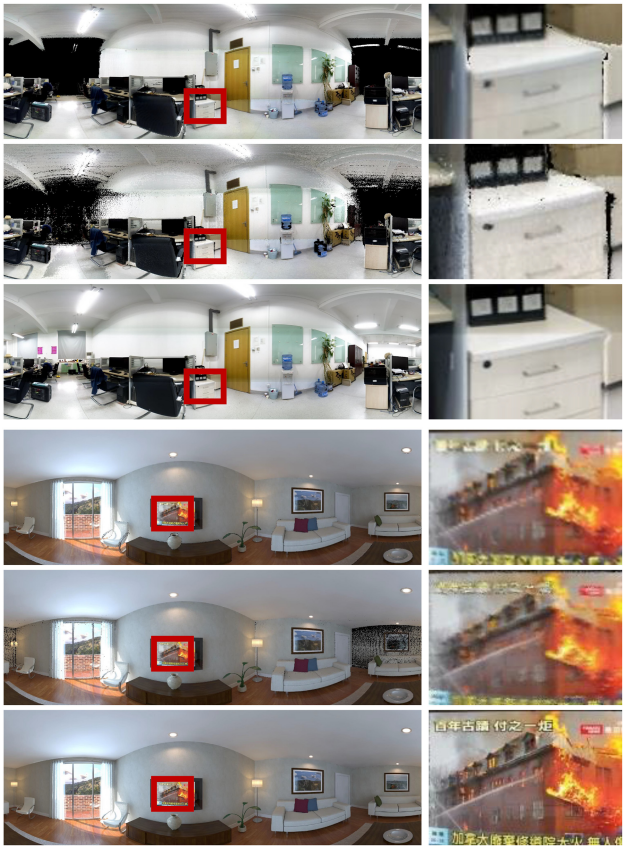


Fig. 3. Comparisons of generated color panoramas using TSDF volumes (the first row for each example), surfels (the second row), and equirectangular images (the third row) with local views on the right. Both TSDF volumes and surfels in this figure use 10-meter cut off but still result in incomplete regions, while the equirectangular projection can correctly project the color information to the panorama domain when given correct relative poses.

conducive to the statistics and optimization of raw depth measurements. Since each raw frame to be integrated only has a small parallax to the constructed panorama, nearly all regions of raw images can be warped into the panorama with little occlusions; thus, such a panorama is able to convey most of the valid measurements. In particular, there are several candidate structures for constructing a panorama, such as a cube map, a stereographic projection image, and an equirectangular image. Among them, the equirectangular image is the best way to evenly reproject raw RGB-D pixels to the target domain, and maintain their neighboring relationship. Hence, it becomes a common choice in both previous methods [Hedman et al. 2017; Hedman and Kopf 2018] and ours.

Figure 3 shows the visual difference between using the TSDF volume, surfels, and the equirectangular image for fusing color measurements, where both TSDF volume and surfels take much more memory (on average, 4 GB for TSDF volume with voxel hashing [Nießner et al. 2013] and 1.8 GB for surfels, respectively), but result in poor quality after ray-casting. As a comparison, the chosen equirectangular image format simply needs 1 GB to store those reprojected depth measurements for statistics.

In the combination step, we firstly warp every RGB-D frame onto the panorama by constructing an organized 3D mesh through adjacent valid depth pixels for reprojection, where equirectangular projection $\mathbf{K}_e(\cdot)$ is used and $\mathbf{T}_u^{-1} \cdot \mathbf{T}_{ci}$ is assigned as the model-view matrix for each frame i from camera c. Adjacent pixels with their depth difference exceeding a threshold $\lambda_a = 0.15$ m are not connected to avoid generating tiny grids with excessive stretches.

After that, we obtain a 4-channel measurement set for each pixel to conclude the final result, and specifically for the "depth" channel, we replace the definition of depth by the radial distance between the obstacle and viewpoint, since there is no focal plane under equirectangular projection. The most straightforward strategy is averaging, which is applied to the color channels in our implementation. For the more critical depth channel, some 2D-to-3D approaches use Markov Random Field (MRF) [Hedman et al. 2017] for deciding the most suitable values from those multi-view stereo algorithms. However, in the 3D-to-3D case, noise is essentially due to imprecise measurements rather than erroneous visual correspondences. Hence, numerical approaches are feasible for computing, instead of choosing distance values, from these valid measurements.

Our proposed numerical approach is inspired by Zach et al. [2007], who proposed to combine the data fidelity considering all valid measurements with an additional Total Variation (TV) term for maintaining the smoothness between adjacent pixels. In detail, we adjust the final *radial distance* z_i of each pixel $\mathcal{P}_i$ represented in the spherical coordinates $\mathbb{S}$ as $\mathcal{P}_i^{\mathbb{S}} = [\phi_i, \theta_i, z_i]^\top$ (for azimuth, inclination, and radius, respectively) on the panorama O to minimize the following energy function:

$$\underset{z}{\operatorname{argmin}} \sum_{\mathcal{P}_i \in O} \left(\sum_{\mathcal{P}'_j \in \mathcal{I}_i} |z_i - z'_j| + \lambda_b \cdot \nabla z_i \right), \tag{5}$$

where the first term is a data fidelity term, and $\mathcal{P}'_j$ is one of the reprojected raw measurements of $\mathcal{P}_i$, with the set containing all measurements of $\mathcal{P}_i$ denoted as $\mathcal{I}_i$. The latter term is a smoothness term: considering that a majority of indoor scene surfaces are flat as an available feature [Furukawa et al. 2009], this term should correctly formulate this feature for the panoramic image domain. We test three types of candidate formulations of ∇z_i as discussed below, with their balancing parameter λ_b further explained with experiments in Section 6.2.

As the first choice, ∇z_i can be defined similarly as the original form of TV [Rudin et al. 1992; Zach et al. 2007], but such a form based on image gradient would cause spherical surface artifacts due to the changed definition of "depth" measurements for equirectangular images. A modification to diminish the flaw is to convert these spherical coordinates $\mathbb{S}$ into cylindrical coordinates with the cylinder radius $\rho_i = z_i \cdot \sin\theta_i$, and use $\nabla\rho$ instead of ∇z to formulate the TV term, but this form is still not appropriate and may result in cylindrical surfaces.

To better exploit the planarity feature, we attempt to formulate it in the Cartesian coordinates $\mathbb{A}$ with the other two choices for the smoothness term. The second choice is derived from Oswald et al. [2012], which minimizes the total surface area constructed by adjacent pixels in the 3D Cartesian space, as follows:

$$\nabla z_i = \| \Delta \mathcal{P}_\phi^{\mathbb{A},-} \times \Delta \mathcal{P}_\theta^{\mathbb{A},-} \|, \tag{6}$$

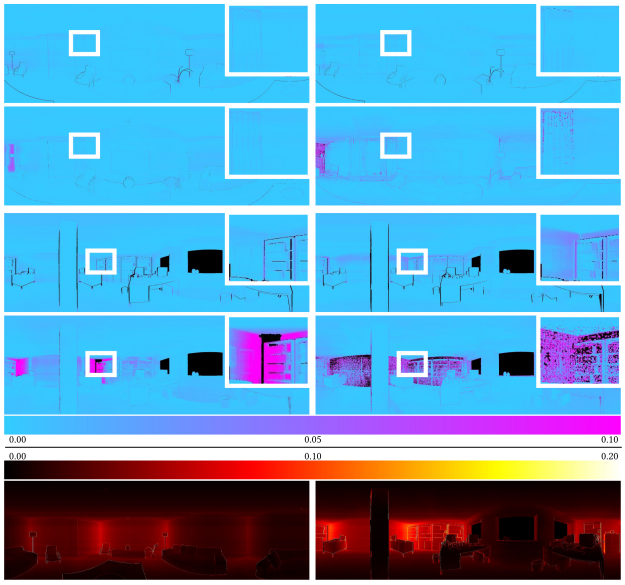


Fig. 4. Comparisons of the geometric quality between different methods, for both the depth value (Top) w.r.t. the ground truth and the uncertainty of depth (Bottom). For the depth comparison (Top)—Left: Equirectangular image domain with the total surface area smoothness term (Equation (6)). Right: Equirectangular image domain with averaging. For the pixel-wise uncertainty after optimization (Bottom)—Left: Ray-casted image from the TSDF volume. Right: Ray-casted image from the surfels. See Table 3 for quantitative differences between three choices of the smoothness term.

where $\Delta\mathcal{P}_x^{\mathbb{A},-} = \overrightarrow{\mathcal{P}_{x-1}^{\mathbb{A}}\mathcal{P}_x^{\mathbb{A}}}, x \in \{\phi, \theta\}$ is the vector formed by adjacent pixels $\mathcal{P}_x$ and $\mathcal{P}_{x-1}$ along two image dimensions ϕ and θ. This term encourages flat regions in the Cartesian coordinates and penalizes uneven surfaces.

The third choice is to measure its total planarity through the normal consistency along two image dimensions, as:

$$\nabla z_i = \|\Delta\mathcal{P}_\phi^{\mathbb{A},-} \times \Delta\mathcal{P}_\phi^{\mathbb{A},+}\|^2 + \|\Delta\mathcal{P}_\theta^{\mathbb{A},-} \times \Delta\mathcal{P}_\theta^{\mathbb{A},+}\|^2, \tag{7}$$

where $\Delta\mathcal{P}_x^{\mathbb{A},+} = \overrightarrow{\mathcal{P}_x^{\mathbb{A}}\mathcal{P}_{x+1}^{\mathbb{A}}}, x \in \{\phi, \theta\}$. We add these squared norms to avoid singularities during iterations when three adjacent pixels are co-linear. This term favors smooth rather than sharp surfaces through two orthogonal surface derivatives.

Figure 4 shows the difference of geometric quality between different data fusion strategies. With the TSDF representation, distant pixels due to a larger uncertainty have higher possibility to erroneously influence their related voxels, and hence are colorized in purple. The surfel representation is able to form dense and normal-consistent surfels for nearby surfaces but the surfels become sparse with their normals disarranged in far regions. Averaging on the equirectangular image outperforms these two data structures for panoramic scanning, and the optimization considering smoothness further enhances the geometric quality of the generated depth maps. Since the performances of the three smoothness terms are close according to our experiments, we refer readers to our evaluation (Table 3) for detailed quantitative comparisons, which shows that the second choice, i.e., the total surface area, slightly outperforms other choices among tested sequences.

Finally, the variance σ_z^2 of the radial distance of such a pixel $\mathcal{P}_i$ is estimated through comparing all its observations $\mathcal{P}_j' \in \mathcal{I}_i$ w.r.t. the solved radial distance, which is then regarded as the uncertainty along the viewing direction of this pixel during subsequent processing in Section 5. On the uncertainty map, we observe that contours as well as distant objects often have higher variance. This phenomenon conforms to the lack of edge sharpness in depth maps and the noise model of raw depth measurements [Handa et al. 2014; Teichman et al. 2013].

5 PANORAMA INTEGRATION

5.1 Noise-aware Alignment between Panoramas

For a fine registration between two panoramas, dense correspondences between their pixels are constructed to formulate and minimize the geometric distance iteratively [Besl and McKay 1992]. Some variations of this strategy further purpose different optimization functions [Lefloch et al. 2017; Rusinkiewicz and Levoy 2001; Segal et al. 2009]. For simplicity, we choose to use the original form of Generalized-Iterative Closest Point (ICP) [Segal et al. 2009] reformed with our obtained pixel-wise uncertainty model, to estimate the relative transformation $\mathbf{T}_{st}$ between two 3D panoramas O_s and O_t:

$$\underset{\mathbf{T}_{st}}{\operatorname{argmin}} \sum_{\mathcal{P}_{si} \in O_s} \|\mathcal{P}_{tj}^{\mathbb{A}} - \mathbf{T}_{st} \cdot \mathcal{P}_{si}^{\mathbb{A}}\|^2_{\Omega^{\mathbb{A}}_{si,tj}}, \tag{8}$$

where $\mathcal{P}_{xi}^{\mathbb{A}} \in O_x$ is the position of a depth pixel i in the 3D Cartesian coordinates $\mathbb{A}$ of panorama O_x. Given a source pixel $\mathcal{P}_{si}^{\mathbb{A}}$, we follow the original nearest neighbor strategy to pick its correspondence $\mathcal{P}_{tj}^{\mathbb{A}}$ on the target frame (again with λ_a for rejecting those exceeding the maximum distance). $\mathbf{T}_{st}$ is initialized based on the estimated transformation between their matched ORB features. Specifically, the covariance $\Omega^{\mathbb{A}}_{si,tj}$ for computing the cost of such a correspondence is calculated as:

$$\begin{aligned} \Omega^{\mathbb{A}}_{si,tj} &= \mathbf{R}_{st} \cdot \Omega^{\mathbb{A}}_{si} \cdot \mathbf{R}_{st}^\top + \Omega^{\mathbb{A}}_{tj}, \\ \text{with } \Omega^{\mathbb{A}}_{xi} &\triangleq \Omega^{\mathbb{A},surf}_{xi} + \Omega^{\mathbb{A},meas}_{xi} \\ &\approx \Omega^{\mathbb{A},surf}_{xi} + \mathbf{J}_{\mathbf{K}_e'} \cdot \Omega^{\mathbb{S},meas}_{xi} \cdot \mathbf{J}_{\mathbf{K}_e'}^\top, \end{aligned} \tag{9}$$

where we use the Gaussian mixture model according to the current estimation of relative rotation $\mathbf{R}_{st}$ for combining covariances of each pair of pixels $\Omega^{\mathbb{A}}_{si}$ and $\Omega^{\mathbb{A}}_{tj}$. For each covariance $\Omega^{\mathbb{A}}$, it is now combined by two parts, namely, the original surface distribution covariance $\Omega^{\mathbb{A},surf}$ [Segal et al. 2009] and our newly considered measurement covariance $\Omega^{\mathbb{A},meas}$. Assuming a normal distribution for each measurement $\mathcal{P}^{\mathbb{S}}$ with covariance $\Omega^{\mathbb{S},meas}$ in the spherical coordinate $\mathbb{S}$, we transform it from $\mathbb{S}$ to $\mathbb{A}$ approximately by the first-order derivative, where $\mathbf{K}_e'(\cdot)$ is the equirectangular back-projection function, and $\mathbf{J}_{\mathbf{K}_e'} \in \mathbb{R}^{3\times3}$ is the Jacobian matrix of $\mathbf{K}_e'(\cdot)$. We assign $\Omega^{\mathbb{S},meas} = \text{diag}(\sigma_\phi^2, \sigma_\theta^2, \sigma_z^2)$, with $\sigma_\phi = \sigma_\theta = 0.5 \cdot \pi/H$ (H being the height of the panorama image) for considering the generated measurement uncertainty during rasterization, and σ_z^2 as summarized during panorama construction.

We use at most 50 iterations for solving each pair of panoramas, as it is sufficient for convergence in experiments. A visualization of the covariance used during the registration is shown

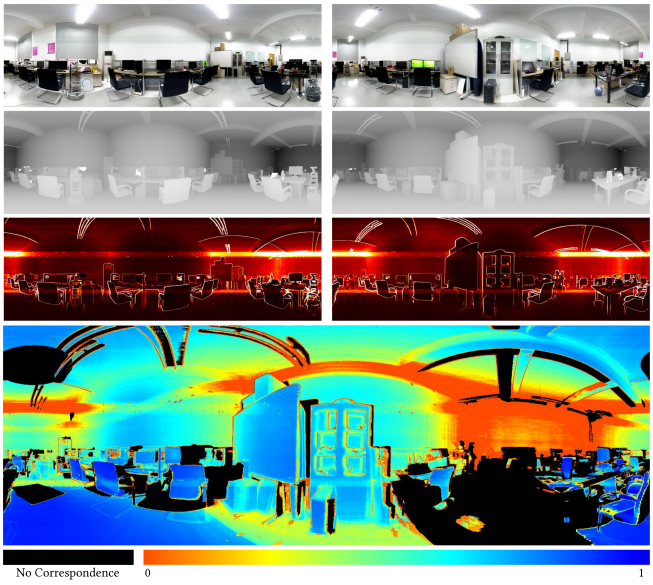


Fig. 5. An example of a registration attempt between two panoramas, matching the source frame (Top-Left) to the reference frame (Top-Right). The location of the source frame in the reference frame is marked as the purple cylinder on the left. Bottom: the weight map of a registration attempt, which is normalized and colorized decreasingly from blue to red. Pixels with no correspondences are drawn in black.

in Figure 5 (bottom). We demonstrate the effectiveness of such a form of mixed covariance, as well as our measurement uncertainty model, by a quantitative comparison to the original form, the combination with a general noise model [Handa et al. 2014], and some other alternative cost functions in Section 6.3.

5.2 Maximum-*a-Posteriori* Integration

To alleviate the inconsistency of surfaces when merging multiple panoramas and generating the final point cloud of the scene, we merge multiple corresponding measurements into one final point during final integration. If a pixel $\mathcal{P}_{tj}$ is the best correspondence of $\mathcal{P}_{si}$ and vice versa, we treat these two pixels as a mergeable pair. Base on the strategy above, we use a union-find algorithm to union those mergeable pairs when associating pixels on all panoramas, and denote the final disjoint set as $\mathcal{U}$.

Finally, if there exists more than one observation $\mathcal{P}_k$ in each disjoint set $\mathcal{U}_j$, we consolidate these measurements through the maximum-*a-posteriori* (MAP) estimation about its final location $\mathcal{Y}_j$ in 3D Cartesian coordinates as follows:

$$\begin{aligned} &\underset{\mathcal{Y}_j}{\operatorname{argmax}} \sum_{\mathcal{P}_k \in \mathcal{U}_j} \mathbf{E}\left(\mathbf{f}_{k,j}^{ob,\mathbb{S}}\right) \\ &\mathbf{f}_{k,j}^{ob,\mathbb{S}} \propto \exp\left(-\frac{1}{2}\|\mathcal{P}_k^{\mathbb{S}} - \mathbf{K}_e(\mathbf{T}_k^{-1} \cdot \mathcal{Y}_j)\|_{\Omega_k^{\mathbb{S}}}^2\right), \end{aligned} \tag{10}$$

with $\mathbf{T}_k$ the final pose of the panorama containing pixel $\mathcal{P}_k$. If a scene contains redundant pairwise registrations, we additionally use a pose graph approach [Grisetti et al. 2010] to refine the final pose of each panorama, with the covariance of each edge set equally during the optimization. For Equation (10), we can deduct an analytical solution for $\mathcal{Y}_j$, as:

$$\mathcal{Y}_j = \left[\sum_{\mathcal{P}_k \in \mathcal{U}_j} (\mathbf{R}_k \Omega_k^{\mathbb{A}} \mathbf{R}_k^{\top})^{-\frac{1}{2}}\right]^{-1} \left[\sum_{\mathcal{P}_k \in \mathcal{U}_j} (\mathbf{R}_k \Omega_k^{\mathbb{A}} \mathbf{R}_k^{\top})^{-\frac{1}{2}} \cdot \mathbf{T}_k \mathcal{P}_k^{\mathbb{A}}\right]. \tag{11}$$

6 EXPERIMENTS AND RESULTS

In this section, we first briefly introduce our data acquisition process and some implementation details as well as the discussions on parameter settings (Section 6.1). We present our evaluation on the quality of panorama construction (Section 6.2) in comparison to previous reconstruction systems for demonstrating the impact of graph factors proposed in Section 4.1, and assess different depth stitching strategies discussed in Section 4.2. The performance of the used noise-aware panorama registration method (Section 5.1) and its subsequent MAP integration (Section 5.2) are also tested with several candidate approaches in Section 6.3. We finally run an experiment to assess the influence of the angular interval of used frames (Section 6.4), and discuss the limitations and possible enhancements (Section 6.5). We also refer readers to our supplementary materials containing reconstructed panoramas and their corresponding point clouds.

6.1 Implementation Details

Real-world assembly. We use the Turtlebot3 as our rotator, which reliably performs in-place rotation and is assembled with an elevated sensor bracket containing three PrimeSense sensors (version 1.08) for separately capturing 480p RGB-D streams at 30 Hz (Figure 1). These sensors are mounted with 45° difference in the inclination angle, with nearly 1° overlap between neighboring cameras and reach almost 135° vertical FoV. In our experiments, we scanned 73 panoramas for different types of scenes including corridors, meeting rooms, offices, and halls. For each panorama, it took the robot about 32 seconds to rotate by 360° in place and produce about 960×3 frames. For each scene, the average distance between panorama locations is about 2 meters. As an example, the office scene shown in Figure 1 (right) uses six panoramic scans for the final integration.

Simulated scans. We prepared simulated scans through the two synthetic scenes presented in Imperial College London and National University of Ireland Maynooth (ICL-NUIM) [Handa et al. 2014]. Robot motions are simulated in Gazebo, a robot simulation platform for experiments, with the trajectories of these sensors extracted for performing highly-realistic rendering. We add depth noise according to the model proposed by Handa et al. [2014]. Synthetic scans are particularly useful for quantitative evaluation due to the available ground truth, and we constructed six simulated scans with five adjacent pairs (within 3.0 meters and able to contain sufficient overlap for registration) on two scenes (denoted as SL for the living room and SO for the office).

System implementation details. Similar to some visual SLAM approaches [Mur-Artal and Tardós 2017], we divide the processing into a front-end for receiving frames and establishing correspondences, and a back-end for continuously performing optimization. Specifically for Equation (4), we choose g2o [Kümmerle et al. 2011] as the framework for solving these optimization prob-

lems. For registering two panoramas, our approach is based on the Generalized-ICP in the Point Cloud Library (PCL) [Rusu and Cousins 2011], which is a single-thread CPU implementation. All experiments were performed on a desktop PC with i7-6850K CPU (3.6 GHz, 6 cores), NVIDIA Titan Xp (12 GB and 3,840 processing units), and 32 GB RAM.

Parameters. Most of the parameters in our system are physically meaningful. Ω^{reg} in Equation (2) reflects the stability of rotation, i.e., adjusted according to the severity of vibration along its 6-DoF as 5.0 mm for translational and 0.5° for rotational standard deviations. Ω^{vel} in Equation (3) is the reciprocal of angular acceleration variance during scanning. According to the consistency of the rotation of the chassis, such variance is assigned as $1^\circ/s^2$. $\lambda_a = 0.15$ m in Section 4.2 and Section 5.1 as discussed before indicates the typical discontinuities between scene instances, and such a threshold is prevalently used in frame registration algorithms [Segal et al. 2009; Whelan et al. 2015a]. λ_b in Equation (5) is assigned according to the chosen smoothness term; see Section 6.2 for details.

6.2 Panorama Construction Quality

Reconstruction with unsynchronized cameras. We first assess the quality of tracking unsynchronized cameras in comparison to various publicly available systems. Since most of them are developed for single camera cases, we pre-stitch the frames from different cameras for subsequent processing, where the result is a sequence of 640 × 1280 images with their focus and focal length the same as the middle camera in order to maximize the used range of observed regions. During stitching, we use two types of relative transformations to reproject and generate these stitched frames: (1) extrinsic parameters (EX) and (2) ground-truth relative poses (GT). The first type is easy to obtain in practice but causes misalignment (Figure 6). The second type ensures the correctness of asynchronous handling for other approaches to make sure they are not affected by the imperfect input.

We choose two sets of algorithms developed for RGB-D scans for comparison: (1) Dense reconstruction methods based on TSDF—InfiniTAM v2 [Kähler et al. 2015] and the state-of-the-art BundleFusion [Dai et al. 2017b]; based on surfels—ElasticFusion [Whelan et al. 2015b]. (2) A representative RGB-D SLAM method: ORB-SLAM2 [Mur-Artal and Tardós 2017]. Both sets are quantitatively evaluated with their original parameters. Since ORB-SLAM2 is not designed for dense reconstruction, we utilize all its keyframes to stitch unorganized point clouds for comparison. To concentrate on the quality of the joint tracking of multiple cameras (Section 4.1) and remove the effects of our proposed panorama integration strategies (Section 4.2), we use both TSDF volume and surfels to fuse these tracked frames in our approach. For our TSDF integration, the voxel size is set as 5.0 mm and truncated by 6.0 cm, as the default configuration suggested by BundleFusion [Dai et al. 2017b]. For surfels, we use the update strategy proposed by ElasticFusion [Whelan et al. 2015b] with their default parameters, and test two versions of depth cut off (3 meters for its default configuration and 4 meters for consistency with BundleFusion). The Root Mean Square Error (RMSE) between the reconstructed models (point clouds or vertices from reconstructed meshes) and ground truth models are calculated by computing the distance of all matching points on these two models.

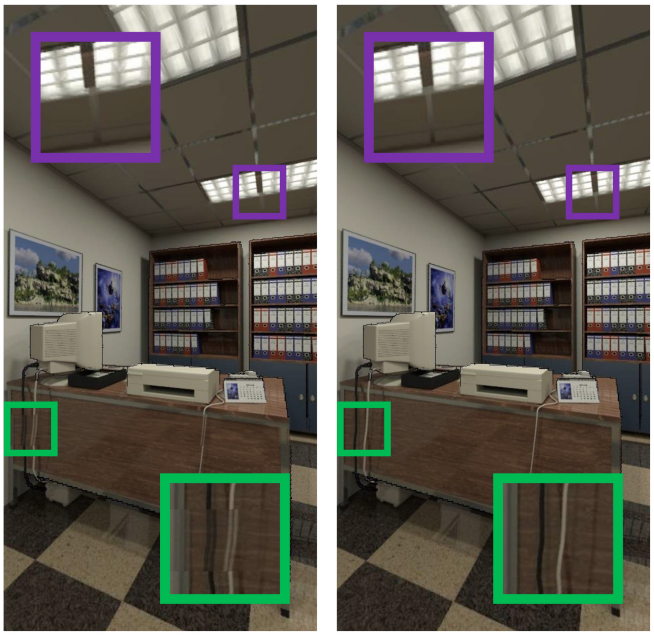


Fig. 6. Left: Pre-stitching three temporally adjacent frames through their extrinsics causes misalignment due to the motion that occurs during the interval between their shutter time (16 ms in this example). The artifacts can be perceived through overlapped areas but are essentially a systematic drifting between frames. Right: Our stitching results according to the tracked pose of these frames.

Table 1. Statistics of Geometric Quality for Different Reconstruction Methods in RMSE (Millimeters)

	SL-1	SL-2	SO-1	SO-2	SO-3	SO-4
InfiniTAM v2 (EX)	116.45	91.32	42.65	67.34	81.97	95.66
ElasticFusion (EX)	89.43	85.61	29.40	49.84	12.36	35.73
BundleFusion (EX)	10.24	26.47	16.23	11.17	19.05	20.09
ORB-SLAM2 (EX)	23.18	33.64	16.44	15.23	17.02	22.36
InfiniTAM v2 (GT)	108.87	81.90	34.71	58.42	80.92	87.35
ElasticFusion (GT)	84.55	79.18	28.98	44.69	10.30	32.46
BundleFusion (GT)	10.20	19.98	16.27	10.46	18.11	17.65
ORB-SLAM2 (GT)	21.06	19.99	16.62	14.36	16.20	21.39
Ours (TSDF Vol.)	11.67	12.03	15.54	17.24	15.42	16.89
Ours (Surfels-4m)	6.87	5.54	5.32	5.34	6.10	6.08
Ours (Surfels-3m)	**5.16**	**4.14**	**4.47**	**4.52**	**3.93**	**4.15**

SX-Y stands for the Y-th scan in the synthetic scene SX. EX stands for using the extrinsics from calibration to stitch these frames, while GT for the ground-truth stitching.

We summarize quantitative comparisons in Table 1. Due to the randomness of the camera start-up time, the average frame interval of pre-stitched pairs among different test cases varies from 0 ms to 16 ms, causing different severity on different scans. As a result, our method with both TSDF volume and surfels for integration achieves better results than other systems on a majority of simulated scans, even when other methods are fed with ground-truth stitched frames, which are hard to acquire in practice. This demon-

Table 2. Reconstruction Quality (RMSE in Millimeters) by Different Approaches with a Single Camera

	SL-1	SL-2	SO-1	SO-2	SO-3	SO-4
Taylor et al.	45.4	167.8	9.4	18.5	16.8	11.9
ORB-SLAM2	46.6	149.3	5.8	20.7	11.1	9.4
Ours (BA only)	46.9	147.8	6.3	19.2	11.5	10.0
Ours	**13.4**	**37.2**	**4.1**	**9.1**	**7.5**	**6.8**

Ours (BA only) is for a comparative experiment, where neither pose regularization factors (Equation (2)) nor velocity smoothness factors (Equation (3)) are used.

Table 3. Statistics of Reconstruction Quality for Different Frame Integration Methods in RMSE (Millimeters)

	SL-1	SL-2	SO-1	SO-2	SO-3	SO-4
TSDF Vol.	11.67	12.03	15.54	17.24	15.42	16.89
Surfels-4m	6.873	5.539	**5.317**	5.343	6.102	**6.076**
Surfels-4m (RC)	7.962	6.921	6.922	6.658	7.628	8.683
ERP Ave.	5.763	4.449	6.778	6.534	7.516	7.401
ERP TV-L^1	**4.915**	3.589	5.596	5.062	6.140	6.245
ERP TSA-L^1	**4.915**	**3.544**	5.515	**5.025**	**6.093**	6.203
ERP TP-L^1	4.986	3.749	5.617	5.080	6.190	6.305

strates our effectiveness of jointly and precisely estimating the trajectories of unsynchronized cameras. Given our tracked poses, the surfel representation outperforms the TSDF volume because these surfels are more flexible, i.e., they need not be fixed at the center of each voxel for constructing the output mesh as vertices. Also, choosing a small cut-off parameter for raw measurements is beneficial for the quality, since the error of depth measurements from such RGB-D cameras is positively related to the distance. However, reducing the maximum distance limits the scope of the reconstruction at each scanning position.

Tracking with a single camera. We next compare our factor graph approach with two SLAM approaches, namely Taylor et al. [2015] and ORB-SLAM2 [Mur-Artal and Tardós 2017], and also with a baseline, which only contains observation factors (Equation (1)). This time, only the middle camera is used since it always contains most feature points for tracking while others are sometimes insufficient. The major difference between our approach and the compared methods are the two additional regularization factors (pose regularization and smooth angular velocity, Equations (2) and (3)) in the graph optimization.

In this experiment, we stitch ground truth depth maps according to the generated trajectory and calculate the RMSE w.r.t. ground truth models, which is thus equivalent to the trajectory assessment, since the severity of the deviation of the trajectory is directly reflected as the quality of the stitched point clouds. From the quantitative results in Table 2, it can be seen that our augmented factors effectively enhance the trajectory by a considerable margin.

Integration with different structures and strategies. We perform experiments on multiple frame integration approaches mentioned in Section 4.2. Seven approaches in total (with three already listed in Table 1) are tested with the same trajectories obtained from the proposed joint tracking of multiple cameras: (1) Integration by TSDF volume, as a general frame integration algorithm. (2) Integration by surfels, with 4-meter adoption to remain consistent with other integration structures. (3) Averaging in the equirectangular image domain. (Actually ERP = EquiRectangular Projection). (4) TV-L^1 performed in the panorama image domain with cylindrical coordinates. (5) Total Surface Area with L^1 (Equations (5) and (6)) denoted as TSA-L^1. (6) Total Planarity with L^1 (Equations (5) and (7)) denoted as TP-L^1. For the balancing parameter λ_b, we traversed its possible value and learned that a good choice should make the ratio of the data term to the smoothness term around 4.0. Hence, we use 50, 5×10^3, and 2×10^8 for TV, TSA, and TP terms, respectively.

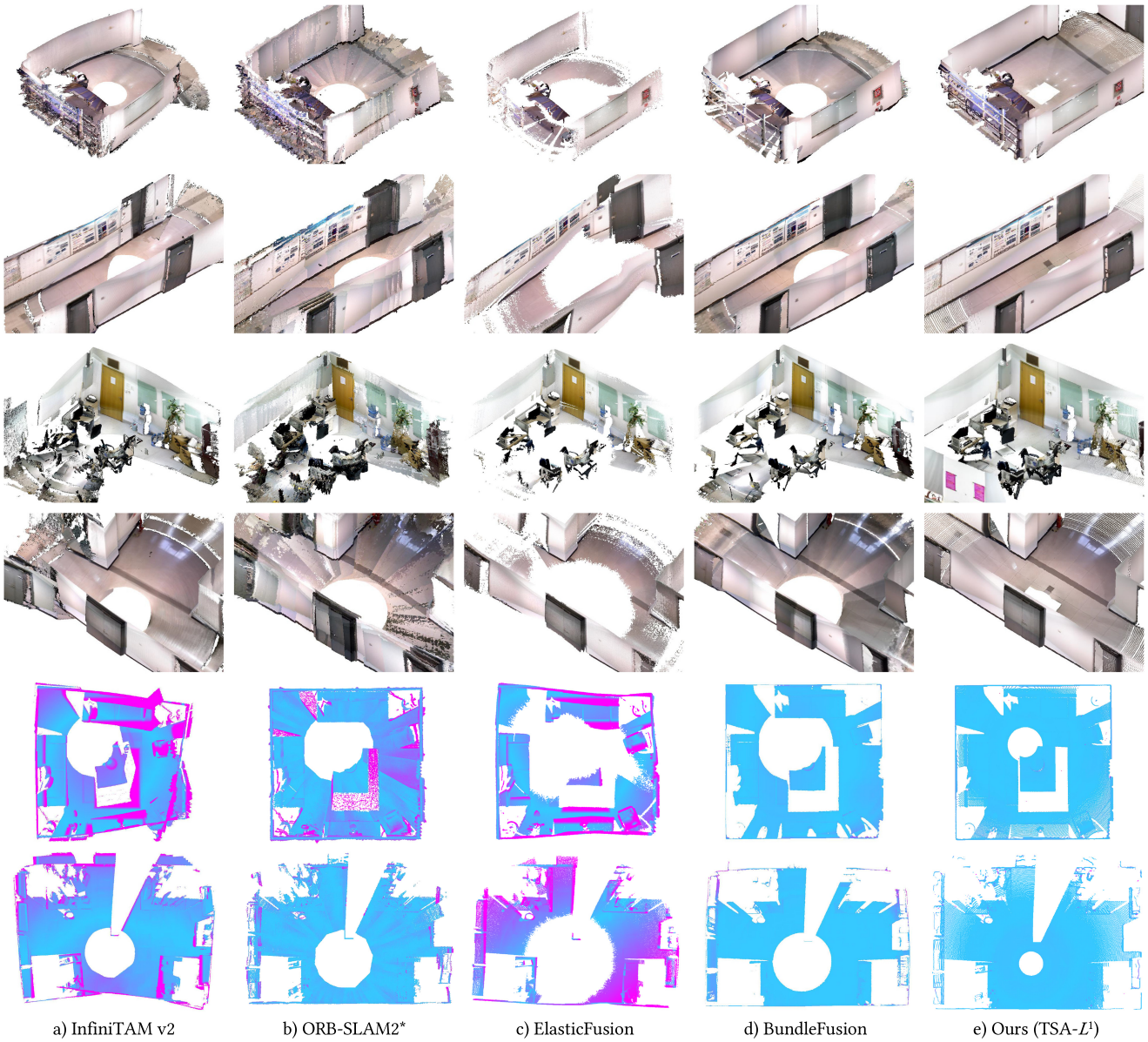
Results of these variants of frame integration are listed with their RMSE w.r.t. the ground truth model in Table 3, which shows the effectiveness of enhancing the geometric quality by presenting optimization in such an image domain rather than the general data structure TSDF. On the other hand, all these three candidate optimization methods achieve better quality than the straightforward averaging, and TSA-L^1 is slightly higher than other methods on all tested panoramas. In addition, this RMSE-based measure is friendly to the surfel-based representation, since such a measure only captures quality statistics of the centroid of each surfel. But when treating surfels as disks for ray-casting images, its quality is also affected by the normal and radius of surfels (see "RC" in Table 3 assessing the correctness of depth pixels on their projected panorama, also with visualized examples in Figure 4).

Visual comparisons on stitched sequences. Figure 7 presents visualized qualitative comparisons between ours (TSA-L^1) and other methods, where the other methods are fed with extrinsically-stitched frames (EX in Table 1). It can be seen that both BundleFusion and ours succeeded to perform robust tracking in various scene types in the test dataset. However, our methods have better reconstruction quality on detailed objects than BundleFusion. In the last two rows, we show bird-eye views and colorize the scenes based on the error of each point/vertex. It shows that our method performs better than BundleFusion, especially in those relatively far areas to the viewpoint.

Additional results for real-world scans are shown in Figure 10. When compared to the synthetic scans, real-world sequences additionally contain more uncertain data located in the intersected region of adjacent cameras (two narrow bands), and some even contain highly uncertain scattered speckles since the depth sensing is mutually interfered by erroneous stereo matching of active IR patterns.

6.3 Registration and Integration for Panoramas

Performance of registration approaches. We demonstrate the effectiveness of our calculated uncertainty model by comparing it to some alternative registration algorithms or configurations. Three categories of cost functions are used for this experiment, as (1) Point-to-point ICP (denoted as p2point) [Besl and McKay 1992], (2) Point-to-plane ICP (denoted as p2plane) [Rusinkiewicz and Levoy 2001], and (3) Generalized-ICP [Segal et al. 2009]. For point-to-plane ICP, we use both the original form, and a variant [Lefloch et al. 2017], denoted as (Curv), which considers both the confidence counter and the curvature for weighting different pixels. For Generalized-ICP, we additionally test its variant, Anisotropic-ICP [Maier-Hein et al. 2012], which chooses to pick the closest point according to the Mahalanobis distance rather than



a) InfiniTAM v2 b) ORB-SLAM2* c) ElasticFusion d) BundleFusion e) Ours (TSA-L^1)

Fig. 7. Results of different reconstruction methods visualized as meshes. For synthetic scenes in the last two rows, meshes are color-coded to show their RMSE. Our approach achieves better geometric quality among these tested scans. For uniform display, we use 4-meter cut off for all output assets in this figure but keep their parameters as default for quantitative evaluation (Table 1).

Euclidean distance. In addition to Generalized-ICP, two alternative forms of covariances are also evaluated, including (1) the original surface uncertainty model (as $\Omega \triangleq \Omega^{surf}$ in Equation (9)), and (2) the mixed uncertainty model (as $\Omega \triangleq \Omega^{surf} + \Omega_0^{meas}$) with a general noise model Ω_0^{meas} derived from Handa et al. [2014] denoted as (Def.). For all tested methods, we remove those depth measurements with their summarized standard deviation σ_z larger than 0.15 meters as unified pre-processing.

Quantitative results are given in Table 4. When compared to the Generalized-ICP, the point-to-plane cost function and its weighted variant are, in fact, a simplification of the general uncertainty-aware form. Hence, its performance is generally worse, but a

Table 4. Registration Quality (RMSE in Millimeters) of Every Pair of Panoramas

	SL-12	SO-13	SO-14	SO-23	SO-24
P2point-ICP	5.918	10.37	6.492	6.528	6.513
P2plane-ICP	4.014	4.870	5.197	5.404	5.951
P2plane-ICP (Curv)	3.914	4.605	4.845	5.162	5.537
Generalized-ICP	3.938	4.533	4.878	4.699	4.650
Anisotropic-ICP	3.957	4.449	4.784	4.558	4.610
Generalized-ICP (Def.)	3.879	4.410	4.736	4.530	4.613
Generalized-ICP (Ours)	**3.867**	**4.399**	**4.722**	**4.523**	**4.594**

SA-XY stands for registering the X-th scan to the Y-th scan of scene SA.

Table 5. Final Integration Quality of Panoramas in RMSE (Millimeters)

	SL-12	SO-13	SO-14	SO-23	SO-24	SO
Align only	3.867	4.399	4.722	4.523	4.594	4.179
Align + Ave.	3.497	4.213	4.495	**4.299**	4.329	3.916
Align + MAP	**3.483**	**4.206**	**4.483**	**4.299**	**4.327**	**3.909**

SO stands for registering all panoramas from the synthetic office scene.

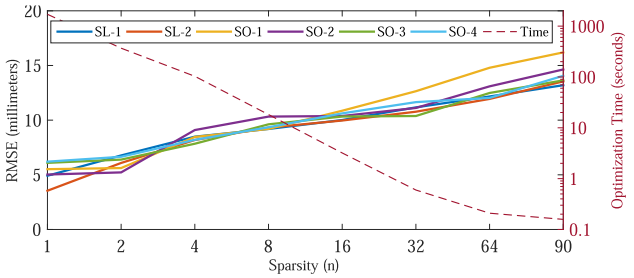


Fig. 8. The impact of the sparsity (controlled by different values of n) on the reconstruction quality and optimization times. The error increases slightly while the time cost is significantly reduced.

weighting scheme emphasizing confidence and low-curvature pixels is applicable. Our proposed strategy, which uses the mixture of the measurement uncertainty and the surface distribution uncertainty, has demonstrated its advantage on all tested scan sequences in comparison to the original form. In addition, through replacing the general measurement uncertainty model (Def.) by our derived variance information, the geometric quality of registration can be further improved. Revising the correspondence searching scheme is beneficial to our quality in theory, but we found only 13.99% of the correspondences are changed by replacing the Euclidean with the Mahalanobis distance [Maier-Hein et al. 2012], leading to limited improvement (0.003 mm on average for RMSE) on these tested registrations.

Performance of the final integration. We further test candidate methods for integrating multiple panoramas after their relative poses are estimated through our proposed inter-panorama registration. Two strategies, namely the Euclidean averaging and the proposed MAP (Equation (10)), are performed for all mergeable groups. Table 5 shows the geometric quality in RMSE among different combinations, which reflects the necessity of merging and the advantage of the MAP integration.

6.4 Sampling Interval for Quality and Efficiency

Finally, we analyze how the the sampling interval of viewing directions for each panorama affects the speed and quality of panorama construction, since when the interval is larger, the scale of the optimization (Equation (4)) will become smaller due to fewer valid observations. In detail, we use one out of every n frames (from 2 to 90, i.e., the horizontal angle interval between adjacent frames from 0.375° to 33.75°) to change the density of involved frames, and record the resulting quality in RMSE with the running time for optimization (Equation (4)) as shown in Figure 8. It can be seen that the reconstruction quality remains at least centimeter-level on the test set, even when most of the frames are skipped. Typically, when n is greater than 16, the optimization cost starts to be

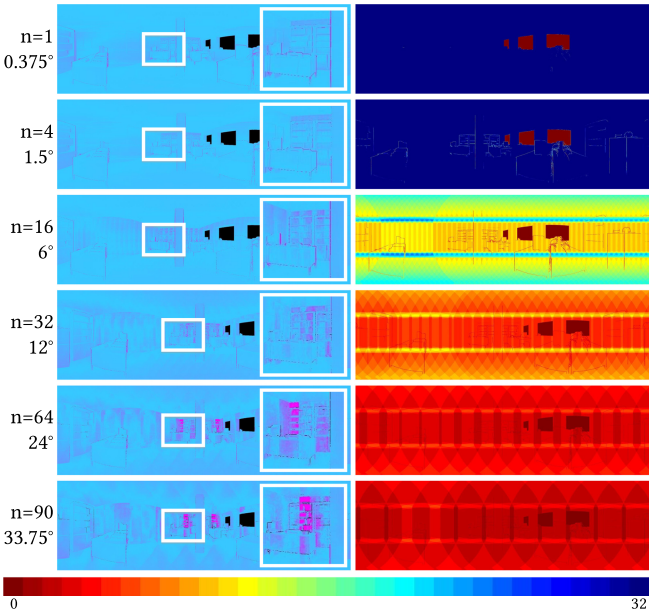


Fig. 9. Results of different sparsity configurations, visualized in depth panoramas colorized by RMSE (Left), and the number of observations (Right) visualized through the colormap (Bottom).

Table 6. Computational Time of Each Module in Our System

Per frame (ms)		Per panorama (s)		Per pano. pair (s)	
Landmarks (front end)	Rendering (Section 4.2)	Tracking (Section 4.1)	Optimize (Section 4.2)	Align (Section 5.1)	Integrate (Section 5.2)
4.3	57.3	3.02	13.94	35.92	2.63

The angular interval of the involved frames (n) mainly affects the time spent for tracking as shown in Figure 9.

acceptable for online applications. The main disadvantage of such acceleration is due to the decreasing number of the observations per pixel, which weakens the reliability of the uncertainty map, especially when the angle interval of adjacent frames exceeds half of their horizontal FoV, resulting in only one observation on some of the pixels (e.g., $n = 64, 90$ in Figure 9). In summary, in real-time applications to support motion planning, we suggest to use about $n = 16$ to balance the quality and efficiency.

Finally, we summarize the average time spent on each operation in Table 6 with $n = 16$ ($60 \times 3 = 180$ frames). The mixture of an additional measurement uncertainty covariance, as well as the analytical solution for MAP integration (Equation (11)), does not bring changes to the time complexity. As a result, our system can perform online reconstruction when the density of involved frames are carefully configured, at the expense of a centimeter-level drop of accuracy. Typically, for a scanning task, it takes about 35 seconds for our platform to perform stable rotation and 30 seconds (with 10 m/s move speed) to go to another viewpoint.

6.5 Limitations

Our proposed approach has several limitations. First, our system does not address dynamic objects, since it is designed to be deployed in fully-static environments. Improvements can be made through integrating a reliable segmentation module [He et al.

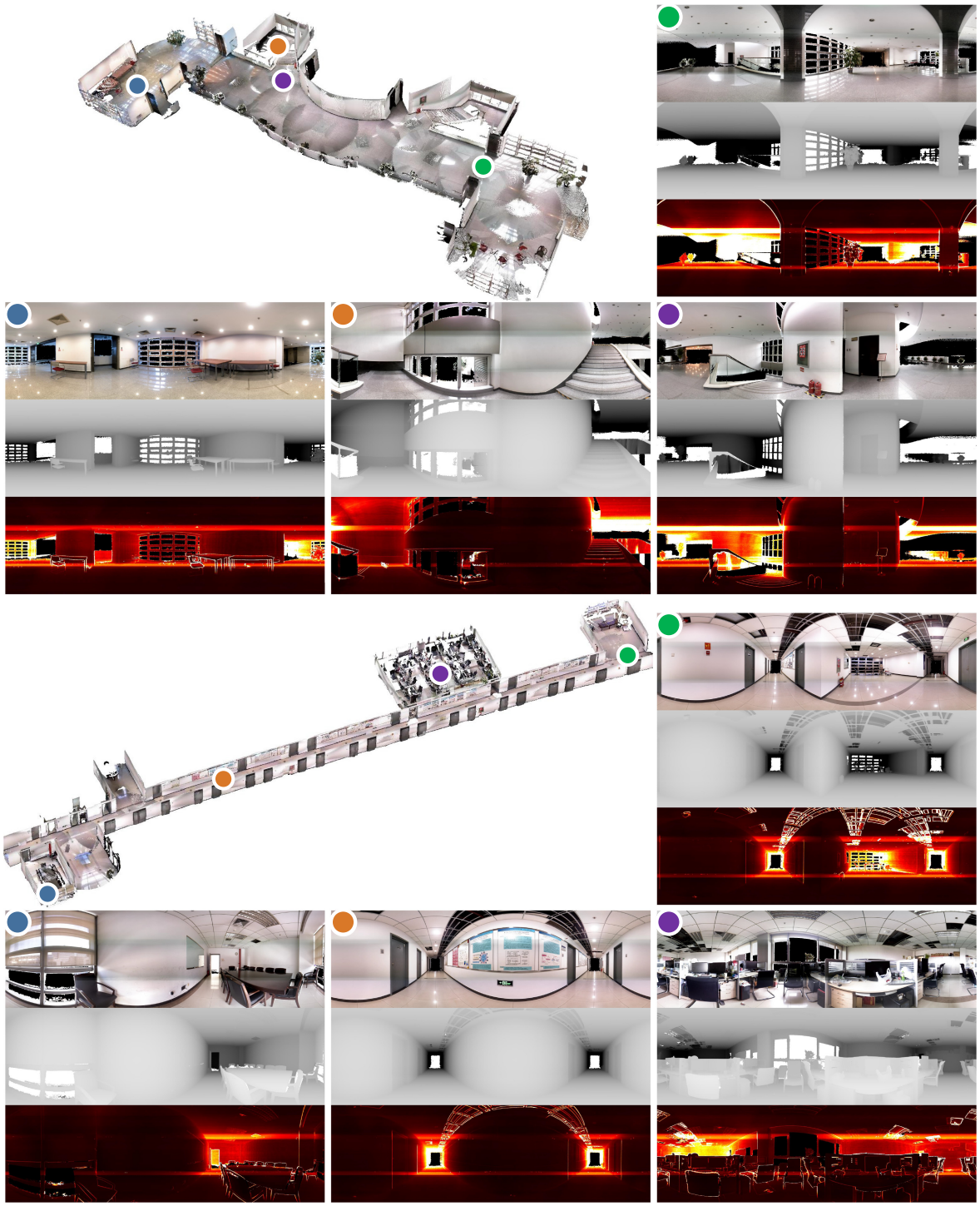


Fig. 10. Panoramas and scenes reconstructed through our proposed algorithm for panoramic scanning.

2017]. Second, color consistency between different cameras is still not satisfied due to their inconsistent exposure and white balance. A post-processing stage for blending colors from different cameras is recommended for a better experience in VR/AR applications. Lastly, depth measurements in overlapped regions have relatively higher noise due to the interference of adjacent cameras. Although our system is able to detect and reduce their influence (as shown in Figure 5 (bottom)), such measurements still require further processing. Scanning on rugged scenes with uneven floor or relatively unstable platforms is also worth testing.

7 CONCLUSION

In this article, we presented a reconstruction system based on a panoramic scanning scheme for successively constructing isolated 3D panoramas and scenes. In the panorama construction stage, we utilize the raw depth information and consensus motion to perform asynchronous camera tracking, and then combine these tracked frames to deduce pixel-wise depth uncertainties, which are subsequently used to provide a high-quality panorama. In the panorama integration stage, multiple panoramas are aligned considering these uncertainties to form the final point cloud of a scanned scene. We demonstrate that our system can be applied to low-cost hardware assembly without additional auxiliary devices such as the time synchronizer or external odometry providers, and succeeds in maintaining sufficient quality for high-fidelity scene representations. In the future, we would like to extend our system to cooperate with a motion planning technique that produces discrete position suggestions to explore and reconstruct indoor scenes autonomously.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their constructive comments.

REFERENCES

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3D semantic parsing of large-scale indoor spaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1534–1543.

Paul J. Besl and Neil D. McKay. 1992. A method for registration of 3-D shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*. 586–607.

C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics* 32, 6 (2016), 1309–1332.

Yan-Pei Cao, Leif Kobbelt, and Shi-Min Hu. 2018. Real-time high-accuracy three-dimensional reconstruction with consumer RGB-D cameras. *ACM Transactions on Graphics* 37, 5 (2018), 171:1–171:16.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*. 667–676.

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. 2015. Robust reconstruction of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5556–5565.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017a. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2432–2443.

Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. 2017b. BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics* 36, 3 (2017), 24:1–24:18.

Wei Dong, Qiuyuan Wang, Xin Wang, and Hongbin Zha. 2018. PSDF Fusion: Probabilistic signed distance function for on-the-fly 3D data fusion and scene reconstruction. In *European Conference on Computer Vision (ECCV)*. 701–717.

Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. 2009. Manhattan-world stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1422–1429.

Ben Glocker, Jamie Shotton, Antonio Criminisi, and Shahram Izadi. 2015. Real-time RGB-D camera relocalization via randomized ferns for keyframe encoding. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 21, 5 (2015), 571–583.

Giorgio Grisetti, Rainer Kummerle, Cyrill Stachniss, and Wolfram Burgard. 2010. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine* 2, 4 (2010), 31–43.

A. Handa, T. Whelan, J. McDonald, and A. J. Davison. 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*. 1524–1531.

Christian Häne, Lionel Heng, Gim Hee Lee, Alexey Sizov, and Marc Pollefeys. 2014. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *International Conference on 3D Vision*. 57–64.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. 2017. Casual 3D photography. *ACM Transactions on Graphics* 36, 6 (2017), 234:1–234:15.

Peter Hedman and Johannes Kopf. 2018. Instant 3D photography. *ACM Transactions on Graphics* 37, 4 (2018), 101:1–101:12.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. 2016. SceneNN: A scene meshes dataset with annotations. In *International Conference on 3D Vision*. 92–101.

Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. 2015. Structured indoor modeling. In *IEEE International Conference on Computer Vision (ICCV)*. 1323–1331.

Sunghoon Im, Hyowon Ha, François Rameau, Hae-Gon Jeon, Gyeongmin Choe, and In So Kweon. 2016. All-around depth from small motion with a spherical panoramic camera. In *European Conference on Computer Vision (ECCV)*. 156–172.

Olaf Kähler, Victor Adrian Prisacariu, Carl Yuheng Ren, Xin Sun, Philip Torr, and David Murray. 2015. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 21, 11 (2015), 1241–1250.

M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb. 2013. Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *International Conference on 3D Vision*. 1–8.

Bryan Klingner, David Martin, and James Roseborough. 2013. Street view motion-from-structure-from-motion. In *IEEE International Conference on Computer Vision (ICCV)*. 953–960.

Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. 2011. G2O: A general framework for graph optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*. 3607–3613.

Yasir Latif, Cesar Cadena, and José Neira. 2013. Robust loop closing over time. In *Robotics: Science and Systems*. 30:1–30:8.

D. Lefloch, M. Kluge, H. Sarbolandi, T. Weyrich, and A. Kolb. 2017. Comprehensive use of curvature for robust and accurate online surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 39, 12 (2017), 2349–2365.

L. Maier-Hein, A. M. Franz, T. R. dos Santos, M. Schmidt, M. Fangerau, H. Meinzer, and J. M. Fitzpatrick. 2012. Convergent iterative closest-point algorithm to accommodate anisotropic and inhomogenous localization error. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 34, 8 (2012), 1520–1532.

Oliver Mattausch, Daniele Panozzo, Claudio Mura, Olga Sorkine-Hornung, and Renato Pajarola. 2014. Object detection and classification from large-scale cluttered indoor scans. *Computer Graphics Forum* 33, 2 (2014), 11–21.

Matterport Inc.2019. The Pro2 Camera. Retrieved from https://matterport.com/.

Jérôme Maye, Paul Furgale, and Roland Siegwart. 2013. Self-supervised calibration for robotic systems. In *IEEE Intelligent Vehicles Symposium (IV)*. 473–480.

Raul Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics* 33, 5 (2017), 1255–1262.

Claudio Mura, Oliver Mattausch, Alberto Jaspe Villanueva, Enrico Gobbetti, and Renato Pajarola. 2014. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers and Graphics* 44 (2014), 20–32.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 127–136.

Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. 2013. Real-time 3D reconstruction at scale using Voxel hashing. *ACM Transactions on Graphics* 32, 6 (2013), 169:1–169:11.

M. R. Oswald, E. Töppe, and D. Cremers. 2012. Fast and globally optimal single view reconstruction of curved objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 534–541.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 77–85.

Bo Ren, Jia-Cheng Wu, Ya-Lei Lv, Ming-Ming Cheng, and Shao-Ping Lu. 2019. Geometry-aware ICP for scene reconstruction from RGB-D camera. *Journal of Computer Science and Technology* 34, 3 (2019), 581–593.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* 60, 1–4 (1992), 259–268.

S. Rusinkiewicz and M. Levoy. 2001. Efficient variants of the ICP algorithm. In *International Conference on 3D Digital Imaging and Modeling*. 145–152.

Radu Bogdan Rusu and Steve Cousins. 2011. 3D is here: Point cloud library (pcl). In *IEEE International Conference on Robotics and Automation (ICRA)*. 1–4.

P. Schmuck and M. Chli. 2017. Multi-UAV collaborative monocular SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*. 3863–3870.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*. 501–518.

Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. 2009. Generalized-ICP. In *Robotics: Science and System*. 21:1–21:8.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics* 25, 3 (2006), 835–846.

Richard Szeliski. 2006. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision* 2, 1 (2006), 1–104.

Camillo J. Taylor, Anthony Cowley, Rafe Kettler, Kai Ninomiya, Mayank Gupta, and Boyang Niu. 2015. Mapping with depth panoramas. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6265–6272.

Alex Teichman, Stephen Miller, and Sebastian Thrun. 2013. Unsupervised intrinsic calibration of depth sensors via SLAM. In *Robotics: Science and System*. 27:1–27:8.

T. Weise, T. Wismer, B. Leibe, and L. Van Gool. 2009. In-hand scanning with online loop closure. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1630–1637.

Thomas Whelan, Michael Kaess, Hordur Johannsson, Maurice Fallon, John J. Leonard, and John Mcdonald. 2015a. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research (IJRR)* 34, 4–5 (2015), 598–626.

T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. 2015b. ElasticFusion: Dense SLAM without a pose graph. In *Robotics: Science and Systems*. 1:1–1:9.

Erik Wijmans and Yasutaka Furukawa. 2017. Exploiting 2d floorplan for building-scale panorama RGBD alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1427–1435.

S. Yang, B. Li, M. Liu, Y. Lai, L. Kobbelt, and S. Hu. 2019. HeteroFusion: Dense scene reconstruction integrating multi-sensors. *IEEE Transactions on Visualization and Computer Graphics* (2019), Preprint. https://ieeexplore.ieee.org/document/8723521.

Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A globally optimal algorithm for robust TV-L1 range image integration. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1–8.

Julio Zaragoza, Tat-Jun Chin, Michael S. Brown, and David Suter. 2013. As-projective-as-possible image stitching with moving DLT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2339–2346.

Yizhong Zhang, Weiwei Xu, Yiying Tong, and Kun Zhou. 2015. Online structure analysis for real-time indoor scene reconstruction. *ACM Transactions on Graphics* 34, 5 (2015), 159:1–159:13.

Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. 2013. Elastic fragments for dense scene reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*. 473–480.

Zhe Zhu, Jiaming Lu, Minxuan Wang, Songhai Zhang, Ralph Martin, Hantao Liu, and Shimin Hu. 2018. A comparative study of algorithms for realtime panoramic video blending. *IEEE Transactions on Image Processing* 27, 6 (2018), 2952–2965.

Received August 2018; revised February 2020; accepted March 2020