IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2020, accepted January 2, 2021, date of publication January 13, 2021, date of current version January 28, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3051257*

# Robust 3D Reconstruction Using HDR-Based SLAM

**CHIA-HUNG YEH[1,2], (Senior Member, IEEE), AND MIN-HUI LIN[1]**
[1]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[2]Department of Electrical Engineering, National Taiwan Normal University, Taipei 10610, Taiwan

Corresponding author: Chia-Hung Yeh (yeh@mail.ee.nsysu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2218-E-110-007-, Grant MOST 109-2218-E-003-002-, Grant MOST 108-2218-E-110-002-, and Grant MOST 108-2218-E-003-002-.

**ABSTRACT** 3D reconstruction is an important topic in the field of the emerging applications such as smart robotics, virtual reality (VR), augmented reality (AR), and autonomous driving. RGB-D simultaneous localization and mapping (SLAM) technique is widely used in the reconstruction process. However, low light and low textured environment often results in insufficient point features and fails the reconstruction. To address this problem, we propose a robust RGB-D SLAM system using high dynamic range (HDR) image information called HDR-based SLAM. The deep learning based HDR generation method is adopted to map a single low dynamic range (LDR) image into a radiance map which is normalized to exclude the influence of exposure time. We retrained the ORB descriptor patch to fit the normalized radiance maps in the feature matching step. The proposed method can improve the quantitative camera trajectory accuracy and qualitative result of geometry reconstruction. Experimental results show that the proposed method has better performance compared to that of the standard range imaging SLAM under challenging low light environment, which helps expand the applicability of 3D reconstruction system.

**INDEX TERMS** 3D reconstruction, feature-based SLAM, high dynamic range (HDR), deep learning, ORB, low light environment.

## I. INTRODUCTION

On account of its wide range of applications, 3D scene reconstruction has become one of the most important and active research topics in the field of computer vision over the past few years. Thanks to the launch of the consumer-grade depth sensor such as Xtion, Realsense, and Kinect, the pixel-wise depth and color information of the objective could be obtained more efficiently and economically compared to monocular and stereo camera.

Many methods are proposed for robust camera tracking and efficient volumetric integration in 3D reconstruction. Visual simultaneous localization and mapping (SLAM) can estimate camera motion and reconstruct a 3D scene simultaneously. There are two kinds of SLAMs: feature-based and direct (dense) slams. Feature-based methods extract a sparse set of points from each frame and match them temporally by their feature descriptors. Because of the sparse feature set, these systems are sensitive to occlusion. Dense SLAM methods employ the entire image to increase the matching robustness and accuracy in camera pose estimation. However, dense methods rely on minimization of pixel-wise photometric error between intensities images, which require high computation power and mostly cannot achieve real-time processing without the aid of a GPU [41]. Although the dense slams have more accurate pose estimation, feature-based approaches have merits in real-time applications with CPU. A recent improvement has been obtained by making use of geometry constraint to extract and match feature points like ORB-SLAM [25]. However, in low light scene, insufficient points are extracted, which may cause wrong matching. We use HDR images to reproduce more details in the images than the conventional one so that we can extract more reliable feature points to match.

The high dynamic range (HDR) imaging is the technique to reproduce a wider range of brightness levels than the low dynamic range (LDR) imaging, which brings higher contrast to the screen, greater color intensity without being oversaturated, and more details in low light images [1]. For LDR imaging, a scene is captured by using single exposure and the brightness levels are only 256 (8-bit unsigned char), which results in overexposed bright regions or underexposed dark

The associate editor coordinating the review of this manuscript and approving it for publication was Heng Wang.

ones. In contrast, HDR imaging uses 32-bit float values per channel to better represent the luminance information similar to the human visual system. HDR images can be obtained through hardware or software. The hardware method uses multiple devices or a special CCD image sensor, which is usually not for commercial purposes [2], [3]. Alternately, the software one is more applicable, which uses common camera to acquire LDR images first and then transform them to HDR images by multi-exposure or the tone mapping techniques. The most common multi-exposure image fusion technique captures several images of the same scene with different exposure times, and then merge them to generate a HDR image [4]–[6]. When the scene is dynamic or being captured hand-held, the misalignment issue and ghosting artefact need to be dealt with [7], [8]. In addition, a HDR image can also be generated by a single LDR image using tone mapping such as histogram-based methods [9], [10] or deep learning [11], [12].

In the field of computer vision, given that HDR imaging can preserve details in both extremely dark and light regions, it has great potential to facilitate various tasks, such as 3D reconstruction [13]–[15], visual simultaneous localization and mapping (visual SLAM) [16], [17], object recognition [18], and image correction [19]. For 3D reconstruction, Meilland *et al.* [13] is the pioneer work focusing on real-time HDR texture mapping. In their visual SLAM system, gamma-based inverse CRF is used to transform RGB images into radiance domain and use them for tracking. Because the system relies on built-in auto exposure (AE), camera transformation and exposure time need to be estimated jointly. Li *et al.* [14] also relies on AE but decouples exposure compensation from tracking. By using the normalized radiance maps that is independent of exposure time, the tracking becomes more robust. Recently, some researches focus on actively controlling the exposure time [16], [17] to improve visual SLAM in HDR environments. Unlike the previous works, which are based on dense-SLAM systems, we propose a feature-based HDR-SLAM, and incorporate it into the 3D reconstruction pipeline to improve the reconstructed results under low light environments. Additionally, Yeh *et al.* [15] also uses normalized radiance maps as inputs during camera tracking but relies on inverse CRF to generate radiance maps from RGB images. Since the calibration of CRF function is device-based, the method can only be reasonably used to reconstruct 3D scenes using sequences captured by calibrated depth camera. Therefore, it is not as applicable as our adopted deep learning-based HDR generation method.

The rest of this paper is organized as follows. Section II reviews the background knowledge and the related work. Section III presents the details of the proposed 3D reconstruction pipeline. Section IV explains the proposed HDR-based SLAM. Experimental results are demonstrated in Sec. V. Finally, concluding remarks are made in Sec. VI.

## II. BACKGROUND REVIEW

In 3D reconstruction, tracking objectives (camera pose estimation) is one of the most important steps in the whole pipeline. The most straightforward method is that only the registration between the current frame and the previous frame are conducted based on either point-to-point or point-to-plane error matrix called frame-to-frame tracking. However, errors would accumulate as the time goes by. Frame-to-model tracking has been widely used in recent reconstruction frameworks to overcome this problem. Frame-to-model tracking establishes a global model, to which latter frames are aligned and thus reduces the temporal error propagation.

RGB-D SLAM can be categorized into two classes: direct methods and feature-based methods. Direct methods extract all the geometry or photometric information to find relative camera pose through minimizing the photometric error while feature-based methods that extract and match features from color images. Kinect-Fusion is the classic work for direct methods that the depth frame is aligned to a global volumetric model and the iterative closest point (ICP) algorithm is used to estimate the camera pose [20]. However, KinectFusion have some limitations in terms of drift error, high computations and small mapping space.

Most of the following researchers have focused on the performance of KinectFusion. Extended KinectFusion is to extend the measurement range by using a rolling reconstruction volume and color fusion [21]. ElasticFusion can reduce tracking drift error and secure global consistency by detecting local and global loop closures [22]. Zhou and Koltun [23] proposed a dense scene reconstruction method by finding points of interest through density function to preserve detailed geometry of object. The experiments results indicate that this method can obtain globally consistent pose estimation for every frame in the scene to reduce alignment errors. Choi *et al.* [24] introduce the global pose optimization on the basis of line processes, which makes the reconstruction pipeline robust against erroneous alignment.

Feature-based SLAM is efficient because only part of information is used compared with direct methods. ORB-SLAM [25] is one of the classic feature-based monocular SLAM methods; It uses sparse ORB features from the input image as well as local bundle adjustment and pose graph optimization to estimate the camera pose. Engelhard *et al.* [26] proposed a hand-held RGB-D SLAM system for indoor mapping, which includes SURF feature extraction and matching, ICP for pose estimation, and pose graph optimization for refining trajectory in the pipeline. ORB-SLAM2 [27] concurrently handle camera tracking, local mapping and loop closing. Trajectory drift can be improved significantly by bundle adjustment and pose graph optimization. BundleFusion proposed by Dai *et al.* [28] combine sparse SIFT features with pose estimation framework including dense photometric and geometric errors to align current frame with keyframes. Endres *et al.*. [29] evaluated the accuracy, robustness and

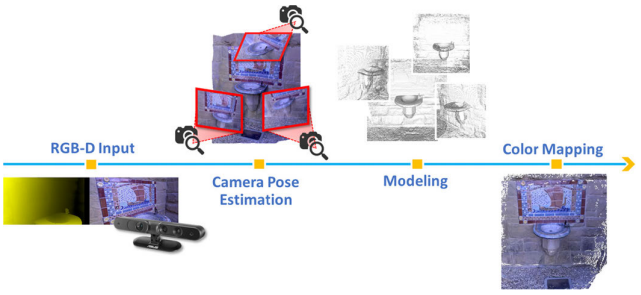


**FIGURE 1.** The pipeline of the 3D reconstruction system.

computations of SIFT, SURF and ORB features, which indicate that ORB is more suitable for the real-time applications.

## III. SYSTEM OVERVIEW

Fig. 1 shows the proposed 3D reconstruction pipeline including the following four steps: (1) Using the consumer-grade RGB-D camera to capture a scene or an object. (2) Employing the proposed HDR-based SLAM to estimate the camera pose. (3) Reconstruct the 3D surface mesh by fusing depth information into the truncated signed distance function (TSDF) volume. (4) Mapping color images onto the geometry reconstruction model. The details of these steps will be elaborated in the following subsections.

### A. CAMERA POSE ESTIMATION

Camera pose estimation plays an essential role in 3D reconstruction because more accurate camera trajectory can generate better geometric model. Camera pose estimation is to localize the camera and contains tracking and optimization. Because visual odometry (VO) estimates the current camera pose through the previous motion state, the measurement errors would accumulate as time goes by and lead to serious odometry drift error. Visual SLAM builds a globally consistent map and uses loop closure detection to detect large loops and correct accumulated drift by pose-graph optimization, so it can produce more accurate camera pose [30]. ORB-SLAM2 is known as one of the well-known Visual SLAM systems, which can perform in real-time on standard CPUs. By integrating loop closing, relocalization, map reuse and bundle adjustment, the reconstruction performance can be improved significantly and applied to a wide variety of environment. Because of lacking the reliable features, ORB SLAM is easy to fail in low light environments. To improve this problem, we proposed HDR-based SLAM which uses HDR images to retain feature matching performance in low light conditions.

### B. 3D SURFACE MESH RECONSTRUCTION

After camera tracking, we integrate RGB-D images and camera poses into a global model by TSDF which is a voxel grid to represent a physical volume of space. The TSDF volume can be regarded as a 3D cube consisting of voxels and each voxel in the volume contains a TSDF value, and the weight. The TSDF value stores the distances from the voxels to the observed surface. Its value is positive when in front of the surface, negative when behind, and nearing zero when at the surface. By using octree data structure to hierarchically partition the TSDF volume and store the TSDF values, the system can handle reconstruction of large-scale scenes given that the octree representation is faster and more memory efficient than the regular grid. Then, marching cube [31] that uses a divide-and-conquer approach to locate the surface in a logical cube is employed to find the zero-crossings in the volume and generate the triangle mesh.

### C. COLOR MAPPING

Color information of the interested 3D object is very important for high-fidelity digitization results, which the surface shape cannot provide by itself. Color mapping is the last step of 3D reconstruction; it maps several color frames to the surface of a model to generate textured results. The goal of color mapping is to estimate the color of each point in the 3D model, and the most straightforward method is to average the values at the corresponding positions of all images. Many researches focus on labeling the corresponding points between the images and the model; however, the process is usually time-consuming especially when the image size is large. Also, it is hard to estimate the accurate point positions on the model surface with no texture. Some works try to align the features in the color images to those in the model; however, the model may not accurate enough, which results in performance degradation. Furthermore, 3D model is always reconstructed through the depth information, and the camera poses of depth maps can be used as reference of their corresponding color images.

In this paper, the comprehensive multi-view stereo texturing methods [32] are used to generate the color texture. Compared with volumetric blending used in many reconstruction systems [33], it can solve the blurring, ghosting and other visual artifacts and generate better results. We will not show the textured models in experiments because our goal is to improve the model reconstruction in low light condition, but the texture would affect the qualitative evaluation of geometry reconstruction.

## IV. PROPOSED HDR-BASED SLAM

Compared to LDR images (the common color images), HDR images in float format can present broader range of luminance in the real environment. The HDRFusion [14] that is based on direct fusion method shows both tracking and mapping can be improved by integrating the radiance map into the SLAM system. Here, we applied the concept on the improvement of the feature-based SLAM methods (ORB-SLAM2), and two modifications are made: (1) Use normalized radiance maps and depth images as input instead of RGB-D images and (2) Train the patch-descriptor especially for normalized radiance maps.

### A. RADIANCE MAP GENERATION

When the depth sensor is held by hand to record a scene or an object in sequence, the common HDR imaging methods combining multiple images is not applicable because they may have different exposure time. In the traditional methods, the camera response function (CRF) can map the relationship between RGB pixel values to radiances, the inverse CRF $f^{-1}$ is used to generate an HDR image from a single exposed LDR image. The CRF is defined as in [34]:

$$B = f\left(R + n_s\left(R\right) + n_c\right) + n_q, \tag{1}$$

where $B$ is a pixel brightness value ranged from 0 to 255 and $R$ is a radiance value, $n_s$ is the noise dependent to radiance, $n_c$ is the constant noise and $n_q$ is the additional quantization noise, which can be ignored. Also, both the means of $n_s$ and $n_c$ are equal to zero, and their variances are defined as:

$$Var\left(n_s\right) = R\sigma_s^2, \tag{2}$$

$$Var(n_c) = \sigma_c^2, \tag{3}$$

Because the CRF of each camera is different, the pre-calibrated process is needed. The calibration process is to set the depth sensor at fixed position, and capture images with different exposure times. By using Debevec *et al.'s* method [4], given the captured images with different exposure times, camera response curve for each color channel can be recovered. Therefore, with the estimated CRF of our depth sensor, and inverse CRF can be calculated directly to transform a single LDR image to a radiance map.

The above methods are primarily model-driven; the requested various camera parameters make it difficult to suit all types of applications. In the recent years, deep learning has led to very good performance on image processing. Here, we include the CNN based method to transfer LDR images to HRD images to increase the feasibility of the proposed 3D reconstruction system. Marnerides *et al.* [35] proposed a new multiscale CNN architecture, called ExpandNet, for high dynamic range expansion from low dynamic range content.

The ExpandNet has three branches in the architecture and they are local, dilation and global branches. In the local branch, network learns the ways to maintain and expand high frequency detail while the dilation branch learns similar information with the larger receptive field. The global branch provides overall information by learning the global context of the resized input. Each branch accepts low light LDR images as input and is responsible for a particular aspect: the local branch handles local detail, the dilation branch handles medium level detail, and the global branch handles higher level image-wide features.

The non-linear transformations of CNN in three branches are given an input vector $x$, so a network of $i$ layers can be expressed as:

$$f_{branch}\left(x\right) = \left(H_i \otimes H_{i-1} \otimes \cdots \otimes H_2 \otimes H_1\right)\left(x\right), \tag{4}$$

where $H_i$ is the $i^{th}$ hidden layer in each branch and $\otimes$ is the composition operator. The output of the fusion layer for HDR image can be written as:

$$f_{fusion}\left(y\right) = \sigma\left[b + W\left(y_{local} \oplus y_{dilation} \oplus y_{global}\right)\right], \tag{5}$$

where $y_{local}$, $y_{dilation}$ and $y_{global}$ are the outputs of the three branches, $W$ is the weight matrix of the fusion layer, $b$ is the bias, $\sigma$ is the scaled exponential linear unit (SELU) activation function and $\oplus$ is the concatenation operation.

Both $W$ and $b$ are learnable part of the fusion layer to concatenate the multi-features at each spatial location, which combines the features to obtain the color and detailed information for the HDR imaging. Then, a small one-layer network called fusion layer processes the extracted features of three branches to transform a single LDR image to a HDR prediction.

### B. NORMALIZED RADIANCE MAP GENERATION

Radiance measures the amount of luminance a sensor captured within exposure time, $\Delta t$, which is formulated as $R = L\Delta t$. In [14], the normalized radiance map is defined as:

$$\begin{aligned}\overline{R_N}\left(u\right) &= \frac{R_N\left(u\right) - E(R_N)}{\sqrt{Var(R_N)}} = \frac{L_N\left(u\right)\Delta t - E(L_N\Delta t)}{\sqrt{Var(L_N\Delta t)}} \\ &= \frac{L_N\left(u\right) - E(L_N)}{\sqrt{Var(L_N)}},\end{aligned} \tag{6}$$

where $N$ is the $80 \times 80$ patch, $u$ is a pixel location in the patch $N$, $\overline{R_N}\left(u\right)$ is the normalized value at pixel $u$, $E(R_N)$ is the mean radiance of the $N$, and $\sqrt{Var(R_N)}$ is the standard deviation of radiances in $N$. For example, for each pixel in a $640 \times 480$ radiance map, and normalization following (6) is performed individually within a $80 \times 80$ window.

Depth sensors have default auto exposure function to better acquire images similar to the one seen by human visual system. When the camera moves from the bright area to the dark area, the exposure time is set longer gradually to make the image brighter. However, if the camera moves fast across the boundary of bright and dark area, the exposure time changes drastically, which results in video flickering. Video flickering would reduce the accuracy of camera tracking, or even fails the tracking. As can be seen in (6), $\overline{R_N}\left(u\right)$ is independent of exposure time $\Delta t$, and this property can fight against video flickers. Because the normalized radiance map is invariant to exposure time, it can better represent the scene than color image does. In addition, HDR images can present wider range of light conditions. Therefore, we use normalized radiance map as input of the proposed HDR-based SLAM system to get better camera poses.

### C. PATCH-DESCRIPTOR TRAINING

HDRFusion is a tacking method directly optimizes the geometry through minimizing photometric errors making use of all the information in the normalized radiance map. In comparison, the proposed HDR-based SLAM system is a featured-based system, which is relatively more efficient as only partial information is involved. In the ORB-SLAM system, the ORB features [36] are extracted from RGB images; then, camera poses calculated by optimizing the projection errors between


**FIGURE 2.** FAST feature detection on color images and corresponding normalized radiance maps.

the features and sparse representation of map is built from the selected features. The feature matching process has three major steps. First, keypoints are detected in an image. Second, feature vectors are used to descript the regions around keypoints. Finally, the corresponding features are obtained by comparing similarities between the descriptors. ORB is a combination of oriented FAST (Features from Accelerated Segment Test) and rotated BRIEF (Binary Robust Independent Elementary Features); it uses oriented FAST for keypoint detection and a rotated BRIEF as the descriptor.

FAST is a corner detection method that compare intensities of the centered pixel with its surrounding circular pixels [37]. We will detect FAST features on normalized radiance map. Fig. 2 shows the FAST feature extraction results of float-format normalized radiance map.

The descriptor, rotated BRIEF, encodes the information around a keypoint into binary strings, so the similarity between two descriptors can be evaluated by calculating their hamming distance. The two features can be regarded as highly-correlated and matched if their hamming distance is smaller than a predefined threshold. To generate the binary descriptor, the patch is centered at a keypoint to include 256 pairs of points. For each patch, if the intensity value of one point in the pair is larger than the other point, the descriptor value would be '1', otherwise it would be '0'. After that, we get a 256 binary string to describe the keypoint.

In the rotated BRIEF, the patch is trained by 300k keypoints in the PASCAL 2006 dataset. The training process is designed to learn 256 pairs from about 200k possible pairs and ensure that they have the two properties, uncorrelation and high variance [36]. Uncorrelation means that the difference between each pair should be as large as possible, thus maximizing the amount of information these 256 pairs carry. High variance means it is more discriminative for a feature, so it can respond to different keypoints. Because the distribution of pairs in LDR images is different from those in the normalized radiance map, we need to retrain the patch for the normalized radiance map. The first step is to collect the raw HDR images online to generate normalized dataset. Then, about 200k FAST features are detected in these normalized radiance maps. Finally, the learning process is re-implemented based on the greedy algorithm [36]. Fig. 3 shows the descriptor patch trained by 200k HDR-keypoints.

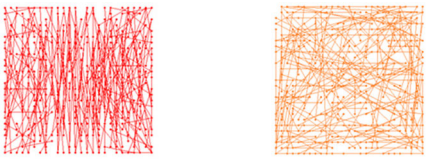


**FIGURE 3.** The descriptor patch trained by keypoints extracted from: (a) color images; and (b) normalized radiance maps.

## V. EXPERIMENTAL RESULTS

To illustrate the robustness and efficiency of the proposed method, we have carried out some experiments on the real-world scene data (TUM RGB-D dataset [38] and our dataset). The quantitative and qualitative comparisons are performed with ORB-SLAM2 which uses LDR images as inputs.

### A. LOW LIGHT REAL-WORLD SCENE DATASETS

In the experiments, two types of datasets are used to evaluate the proposed method: TUM RGB-D dataset and our dataset. The TUM RGB-D dataset consists of calibrated color and depth sequences recorded with full frame rate (30 FPS) using Microsoft Kinect sensor, and provides the ground truth of camera trajectories obtained from a high-accuracy motion-capture system. The other dataset is recorded by us with the other kind of depth camera: Asus Xtion, which can provide the testing datasets with a wide variety of consumer depth cameras. In total, four sequences are selected for evaluation, including three sequences from TUM RGB-D dataset and one sequence from our dataset. All testing datasets are handheld sequences and their detailed information is shown in Table 1: fr1_xyz captures an office desk and contains primarily translation motions along the principal axes of the depth camera; fr1_360 scans from the office center and makes a 360-degree turn with fast wave motion.; fr1_room moves around and scans the whole office; cafeteria captures a long scene consists of potted plants and dining tables, and flickering happens when the camera moves fast across bright and dark regions.

To test the robustness of the proposed method, we further augment the real-world scene datasets to more challenging low light environment. Low light LDR images are simulated using the function EnhanceBrightness in open-source image processing library imgaug [39], which creates an augmenter that reduces the brightness of an image by a selected factor

**FIGURE 4.** Generated low light LDR images of dataset fr1_xyz and cafeteria.

**TABLE 1.** Detailed information of the testing datasets.

| Dataset | Sequence | Duration (s) | Length (m) | Frames |
|---|---|---|---|---|
| TUM RGB-D | fr1_xyz | 30.09 | 7.112 | 790 |
| TUM RGB-D | fr1_360 | 28.69 | 5.818 | 744 |
| TUM RGB-D | fr1_room | 48.90 | 15.989 | 1352 |
| Our dataset | cafeteria | 22.37 | 9.299 | 678 |

ranging from 0 to 1. Here, we set the brightness factor at 0.1 to generate low light real-world scene datasets. Fig. 4 shows the generated low light LDR images of dataset fr1_xyz and cafeteria.

### B. EVALUATION PROCESSES

In the experimental results, we provide quantitative evaluation of camera tracking (section V-C) and qualitative evaluation of 3D geometry reconstruction (section V-D) using low light testing datasets on three methods: (a) ORB-SLAM2 with LDR inputs; (b) HDR-SLAM with HDR inputs; and (c) HDR-SLAM with normalized HDR inputs. For quantitative evaluation, there are two reasons why we only focus on camera tracking: first, ground truth camera trajectories are easier to obtain than ground truth 3D geometry models; second, when estimated camera trajectory is more accurate, the quality of the final reconstruction result is usually better. Therefore, quantitative evaluation metric absolute trajectory error (ATE) is used, and tracking information including percentage of losing tracking, average number of keypoints, and average number of matched map points are also presented in section V-C.

ATE is frequently used for evaluating tracking accuracies in visual SLAM systems by measuring the absolute pose differences between the estimated camera poses $P$ and the ground truth trajectory $Q$. As shown (7) and (8), the rigid transformation $S$ between Trajectory $P$ and $Q$ is calculated by the least squares method [40], and then root mean squared error (RMSE) is applied to all the translation components of error matrices $E$ at each time $i$ to get the $ATE_{RMSE}$.

$$E_i = Q_i^{-1} S P_i, \tag{7}$$

$$ATE_{RMSE} = \left( \frac{1}{n} \sum_{i=1}^{n} \| trans(E_i) \|^2 \right)^{\frac{1}{2}}, \tag{8}$$

In addition to $ATE_{RMSE}$ which measures the accuracy of the estimated camera poses, we also show other tracking information that can provide more insights into the whole camera tracking process in section V-C, including percentage of losing tracking, average number of keypoints, and average number of matched map points. The meaning of these tracking information are: when the percentage of losing tracking is high, fewer aspects of depth information are fused into the TSDF volume and would lead to incomplete reconstructions; etection of representative keypoints is the foundation of feature-based visual SLAM system; because matches between current feature points and existing local map points are essential to minimize the reprojection errors in camera pose estimation step, representative matched map points can ensure more accurate and drift-free localization.

**TABLE 2.** Comparison of tracking information during tracking on low light testing datasets.

| Sequence | Evaluation Metrics | ORB-SLAM2 LDR | HDR-SLAM HDR | HDR-SLAM Normalized HDR |
|---|---|---|---|---|
| fr1_xyz | percentage of losing tracking | 10.89% | **0** | **0** |
| | avg. num. of keypoints | 472 | 992 | **1005** |
| | avg. num. of matched map points | 193 | **292** | 251 |
| fr1_360 | percentage of losing tracking | 81.85% | **0** | **0** |
| | avg. num. of keypoints | 336 | 995 | **1004** |
| | avg. num. of matched map points | 30 | 204 | **212** |
| fr1_room | percentage of losing tracking | 13.09% | **0** | **0** |
| | avg. num. of keypoints | 756 | 1970 | **2004** |
| | avg. num. of matched map points | 181 | 351 | **358** |
| cafeteria | percentage of losing tracking | 52.65% | **0** | **0** |
| | avg. num. of keypoints | 566 | **1005** | 1004 |
| | avg. num. of matched map points | 121 | 158 | **175** |

With the proposed 3D reconstruction pipeline, textured 3D models would be reconstructed. In section V-D, we would present the qualitative evaluation of 3D geometry reconstruction without texturing given that textures would make it hard for us to observe the minor variations in the reconstructed geometry models.

### C. CAMERA TRACKING

First, we show details of the tracking process with percentage of frames lose tracking, average number of keypoints, and average number of matched map points. Second, to verify the improvement of the proposed method in terms of camera tracking accuracy, we evaluate on low light testing datasets using evaluation metric $ATE_{RMSE}$. Here we compare three methods: (a) ORB-SLAM2 with LDR inputs; (b) HDR-SLAM with HDR inputs; and (c) HDR-SLAM with normalized HDR inputs.

Table 2 provides the insights of camera tracking processes. ORB-SLAM2 with LDR inputs has relatively fewer numbers of keypoints and matched map points and therefore would lead to different percentage of failure that lose tracking frames. In contrast, after transforming the LDR images to HDR images and normalized ones, details and contrasts of the images are enhanced, so the number of representative features is significantly increased to improve the tracking robustness. Here, average number of keypoints and matched map points provide insights of the camera tracking process rather than final tracking accuracy. Take dataset fr1_xyz for example, though the average number of the matched map points of HDR-SLAM with HDR inputs is the largest, we will show that, in terms of the stability of matched map points, camera trajectory accuracy, and reconstruction quality, HDR-SLAM with normalized HDR inputs performs the best for all of them.

We also plot the tracking information in time sequence. In Fig. 5 and 6, red background indicates the time when the failure tracking lost cases happen, orange line labels the number of the matched map points, and green line labels the number of detected keypoints. Fig. 5 shows the camera tracking process of each method on low light fr1_xyz dataset and demonstrates two points: (1) ORB-SLAM2 easily loses tracking because of the unstable keypoints and matched points;

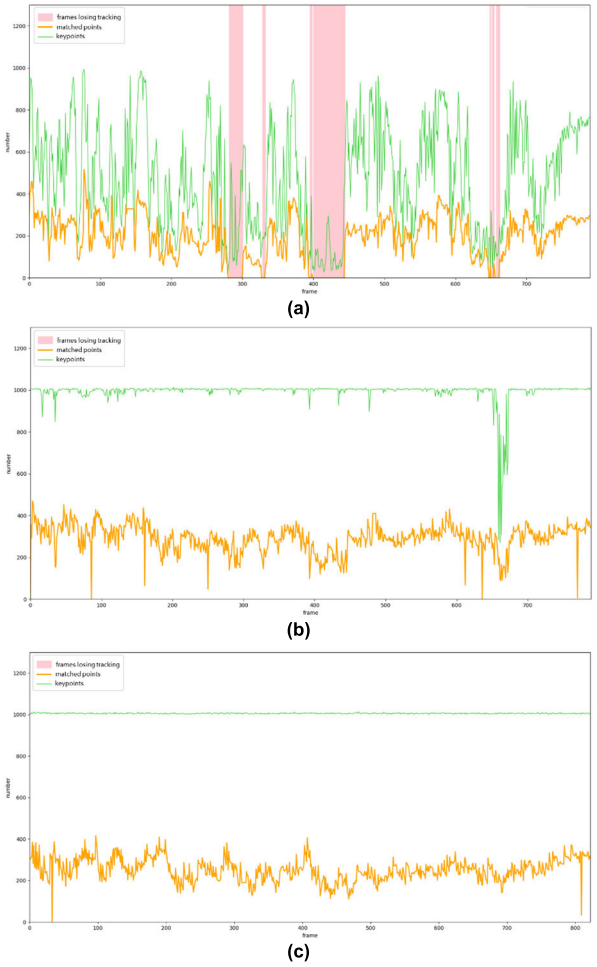


**FIGURE 5.** Camera tracking information of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light fr1_xyz dataset.

(2) for HDR-SLAM with normalized HDR inputs, even though the average number of the matched map points (251) is smaller than the one with HDR inputs (292), the orange curve (matched map points) and the green curve (detected keypoints) are both more stable. Consequently, we can conclude that the matched map points and detected keypoints of HDR-SLAM with normalized HDR inputs has superior

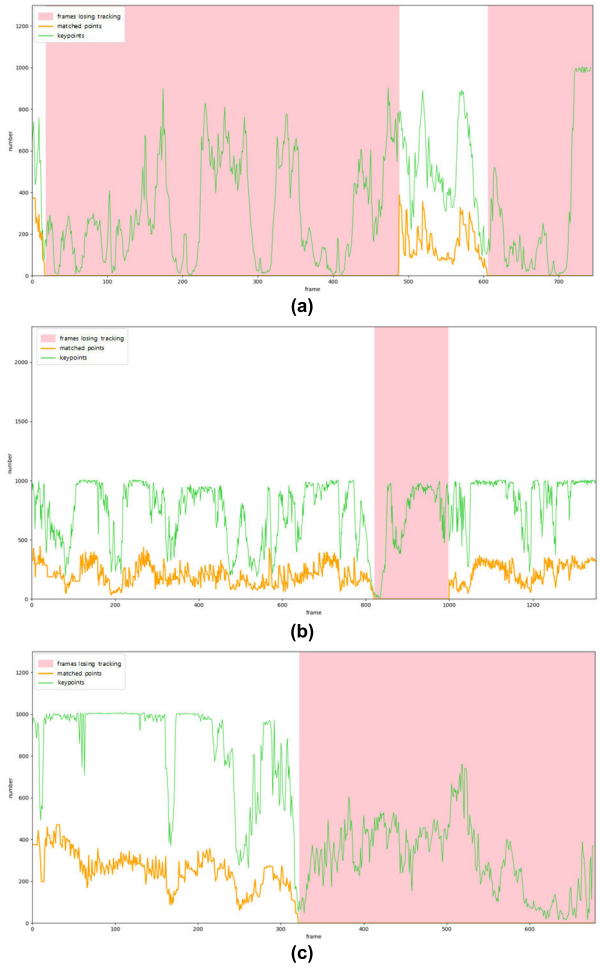


**FIGURE 6.** Camera tracking information of ORB-SLAM2 on low light: (a) fr1_360; (b) fr1_room; and (c) cafeteria testing datasets.

quality and is more invariant to low light environment, low texture, and fast motion.

Fig. 6 shows the camera tracking process of ORB-SLAM2 on other testing datasets. It illustrates that when matched map points decrease rapidly to a small number, failure tracking lost would happen until previously tracked scene is seem again. In ORB-SLAM2, small number of matched map points is prone to occur because of the lack of representative keypoints in low-contrast LDR images. So, in the low light environment, when the captured scene is textureless or the motion is fast or not continuous, ORB-SLAM2 would easily lose tracking.

Table 3 shows the comparison of camera trajectory accuracy using the evaluation metric $ATE_{RMSE}$. Experimental results show that in the challenging low light scenes, ORB-SLAM2 lose tracking in all testing datasets whereas HDR-based methods can successfully track each sequence. Furthermore, since the proposed HDR-SLAM with normalized HDR inputs achieves the best camera trajectory accuracy $ATE_{RMSE}$ on every testing dataset, it can prove that, for HDR-SLAM systems, the normalization of HDR inputs can enhance the robustness of camera tracking. One thing worth

**TABLE 3.** Comparison of camera trajectory accuracy ($ATE_{RMSE}$ in centimeters) on the low light testing datasets.

| Sequence | ORB-SLAM2 | HDR-SLAM | HDR-SLAM |
|---|---|---|---|
| | LDR | HDR | Normalized HDR |
| fr1_xyz | X | 2.055 | **1.977** |
| fr1_360 | X | 27.52 | **20.72** |
| fr1_room | X | 27.17 | **13.93** |
| cafeteria | X | - | - |

X denotes losing tracking; - denotes that all the frames are successfully tracked but $ATE_{RMSE}$ cannot be computed due to the lack of ground truth data.

**TABLE 4.** Comparison of average running speed (FPS) during camera tracking.

| | ORB-SLAM2 | HDR-SLAM | HDR-SLAM |
|---|---|---|---|
| | LDR | HDR | Normalized HDR |
| Average Running Speed | 24 | 16 | 15 |

noting is that, for our dataset cafeteria, all the frames are successfully tracked, but $ATE_{RMSE}$ cannot be computed due to the lack of ground truth camera trajectory.

Fig. 7 (a) illustrates how the ORB-SLAM2 with LDR inputs loses tracking because of the lack of keypoints and matched map points at frame 401-445 on low light fr1_xyz dataset. In comparison, the corresponding frames of HDR-SLAM methods have more representative features and can successfully track each frame. Overall, we can conclude that the proposed HDR-SLAM with normalized HDR inputs can achieve the best performance during the camera tracking stage.

In terms of computation cost, we provide the average running speed (fps) of these three SLAM systems in Table 4. The experiments are carried out with an Intel Core i7-4790 CPU (four cores @ 3.6 GHz). Because feature generation step applied on float-format HDR images is more complex, HDR-SLAMs need more computation power.

### D. 3D GEOMETRY RECONSTRUCTION

#### 1) METHODS BASED ON FEATURE-BASED SLAM

This section demonstrates that the proposed HDR-SLAM with normalized HDR inputs not only estimates the most accurate camera poses, but also generates the best 3D geometry reconstruction results among all the methods. Fig. 8-11 shows the reconstruction results of the three methods on four datasets.

Fig. 8 (low light fr1_xyz): for ORB-SLAM2, due to tracking lost, there are some low-quality and missing areas (marked by red circles); for HDR-SLAM with HDR inputs ($ATE_{RMSE} = 2.055$) and normalized HDR inputs $ATE_{RMSE} = 1.977$), slightly worser camera trajectory accuracy does not cause obvious defection on geometry model because the office desk is scanned back and forth repeatedly; the results of the two HDR-SLAM method are both good.

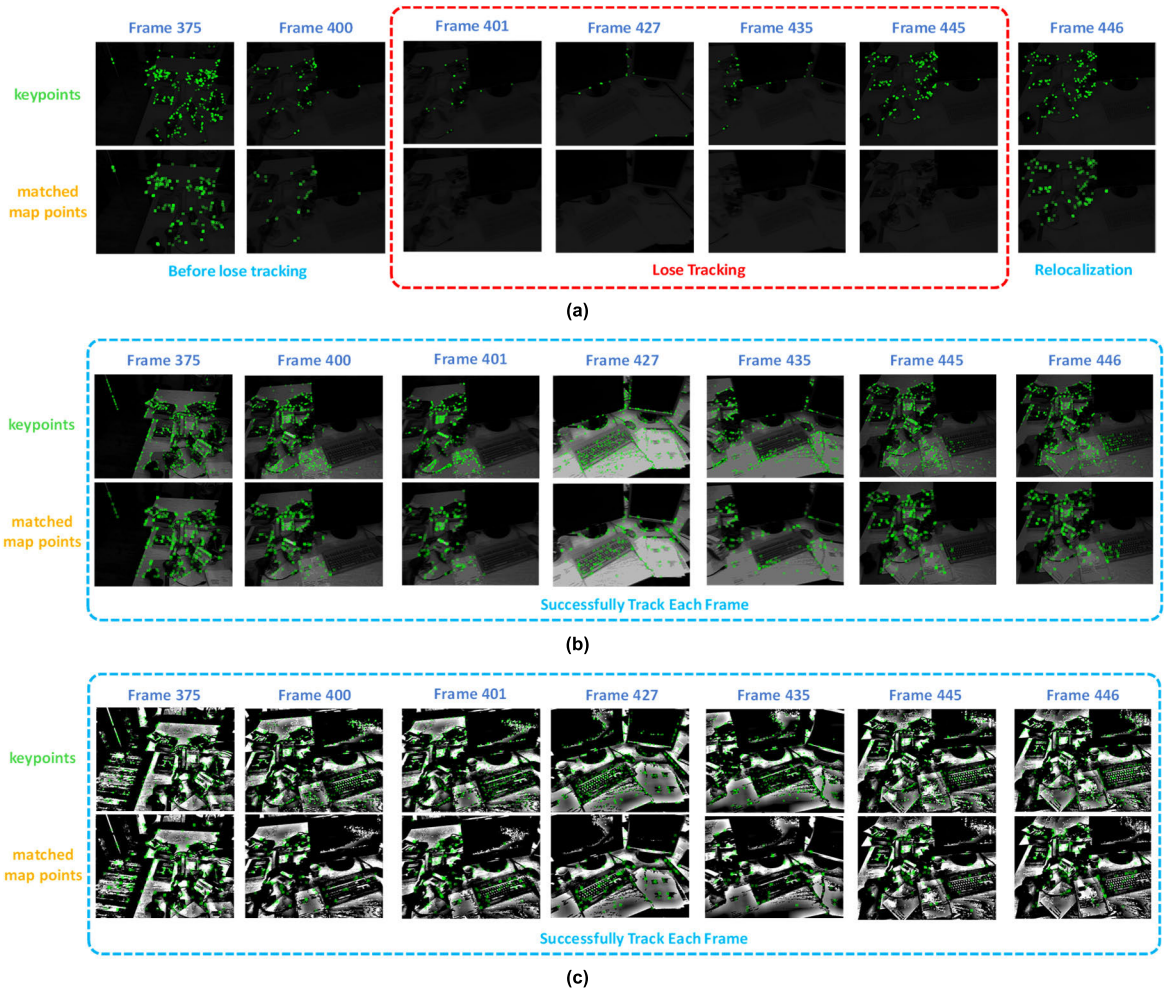


**FIGURE 7.** Keypoints detection examples of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light fr1_xyz dataset.

Fig. 9 (low light fr1_360): for ORB-SLAM2, because the percentage of losing tracking is high (81.85%, frame 18-487 and 606-744), only limited part of the office is reconstructed as marked by the blue rectangles; reconstruction quality of normalized HDR inputs around textureless whiteboard region is obviously better than HDR inputs as marked by red circles; fr1_360 is a very challenging sequence because the camera moves very fast and it captures some textureless whiteboard and ground regions, hence even though the proposed method can track all the frames, the calculated camera poses are not precise enough to construct extremely fine 3D model.

Fig. 10 (low light fr1_room): for ORB-SLAM2, because there is a clip of sequence lose tracking (13.09%, frame 821-996 and 998), the areas around the cabinet is not reconstructed as marked by blue rectangles; HDR-SLAM with the HDR inputs suffers from more severe drifting problem as marked by orange circles; reconstruction quality of normalized HDR inputs around the cabinet is obviously better than HDR inputs as marked by red circles.

Fig. 11 (low light cafeteria): for ORB-SLAM2, because it starts to lose tracking in the middle of the sequence and never relocalize (52.65%, frame 322-678), only the areas scanned at first have been reconstructed as marked by blue rectangles; reconstruction quality of normalized HDR inputs of the whole scene is obviously better than HDR inputs as marked by red circles and rectangles.

Overall, for ORB-SLAM2, because losing tracking problem happens, different degrees of degradation in reconstructed models occur, including low-quality, misaligned and missing issues. Additionally, HDR-SLAM with the normalized HDR inputs performs better or equally well in comparisons to the one with HDR inputs. To sum up, the proposed HDR-SLAM with the normalized HDR inputs also performs the best in terms of 3D geometry reconstruction quality.

### 2) DENSE SLAM METHOD

Given that HDRFusion [14] is the first work to incorporate normalized HDR map into the dense RGB-D dense system, we also show the comparison of their reconstruction result. The experiment is built with the source code provided by the authors and an additional NVIDIA GeForce GTX 1080 GPU (8 GB GDDR5X memory) is used. We have followed the authors' instruction to calibrate our Xtion depth camera and set up parameters to enable the reconstruction on our dataset:

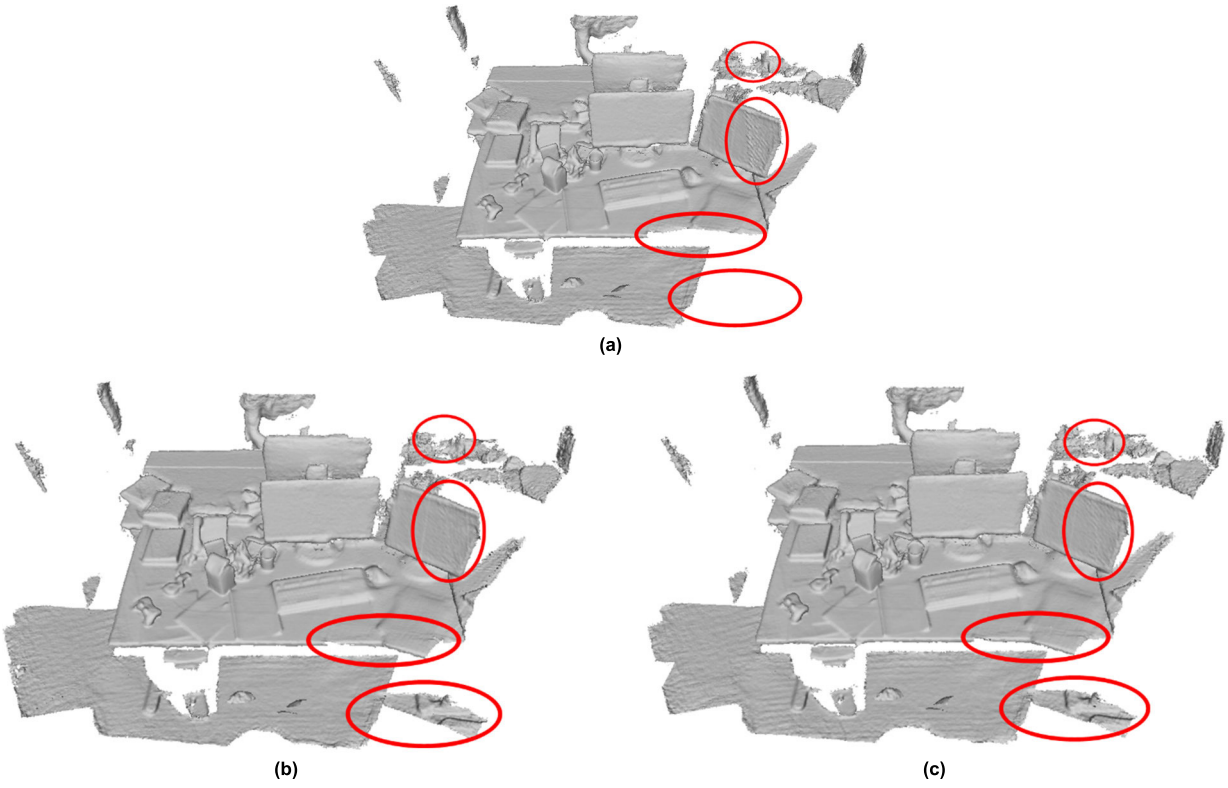


**FIGURE 8.** 3D geometry reconstruction results of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light fr1_xyz dataset.

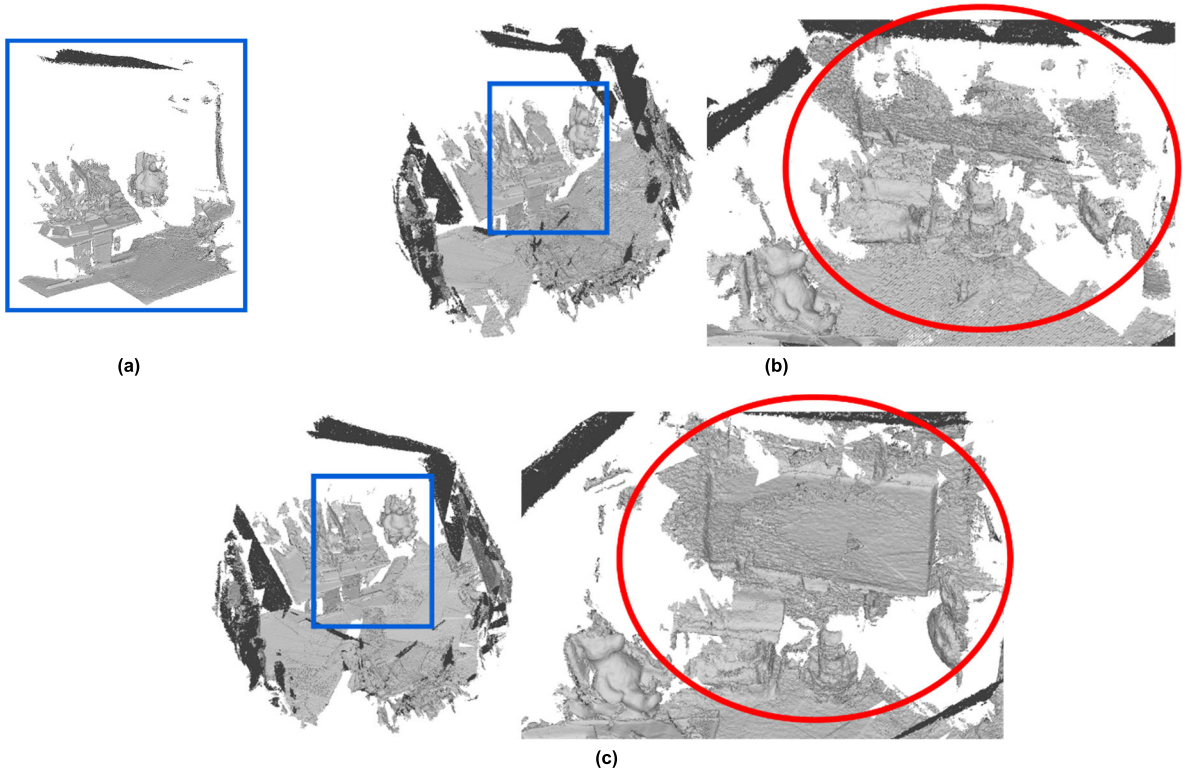


**FIGURE 9.** 3D geometry reconstruction results of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light fr1_360 dataset.

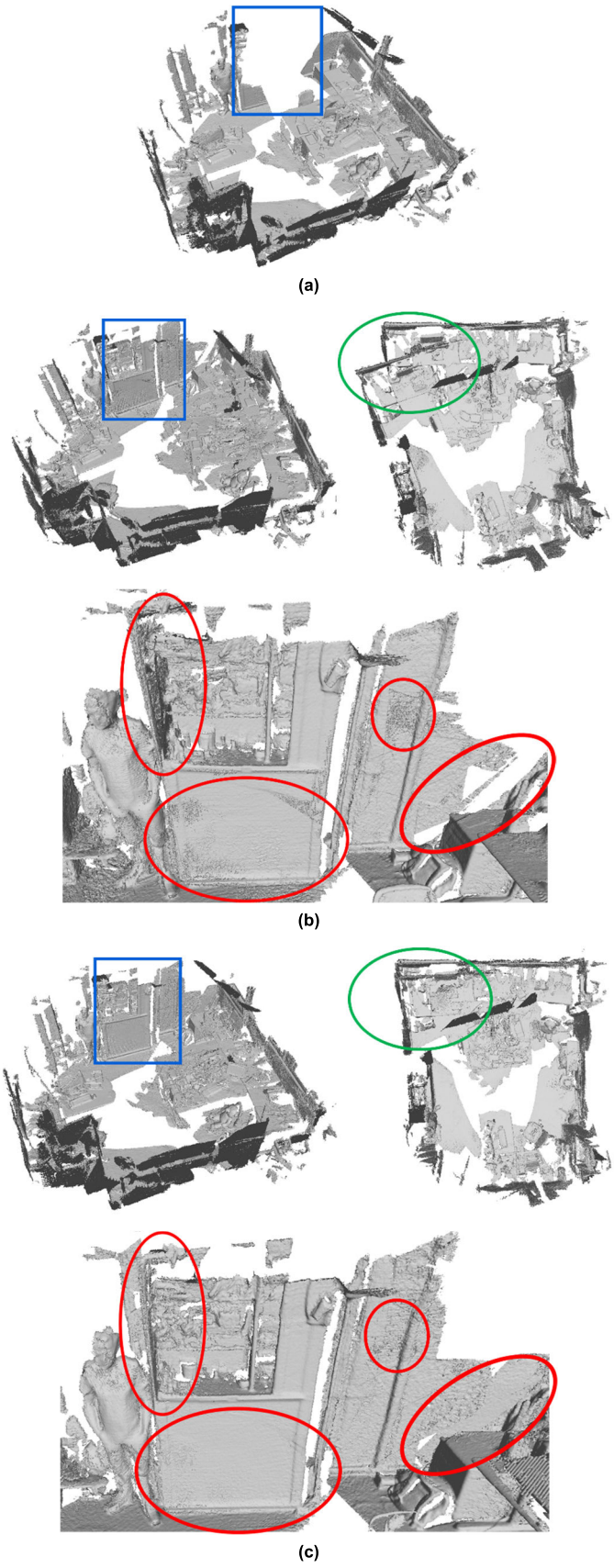


**FIGURE 10.** 3D geometry reconstruction results of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light fr1_room dataset.

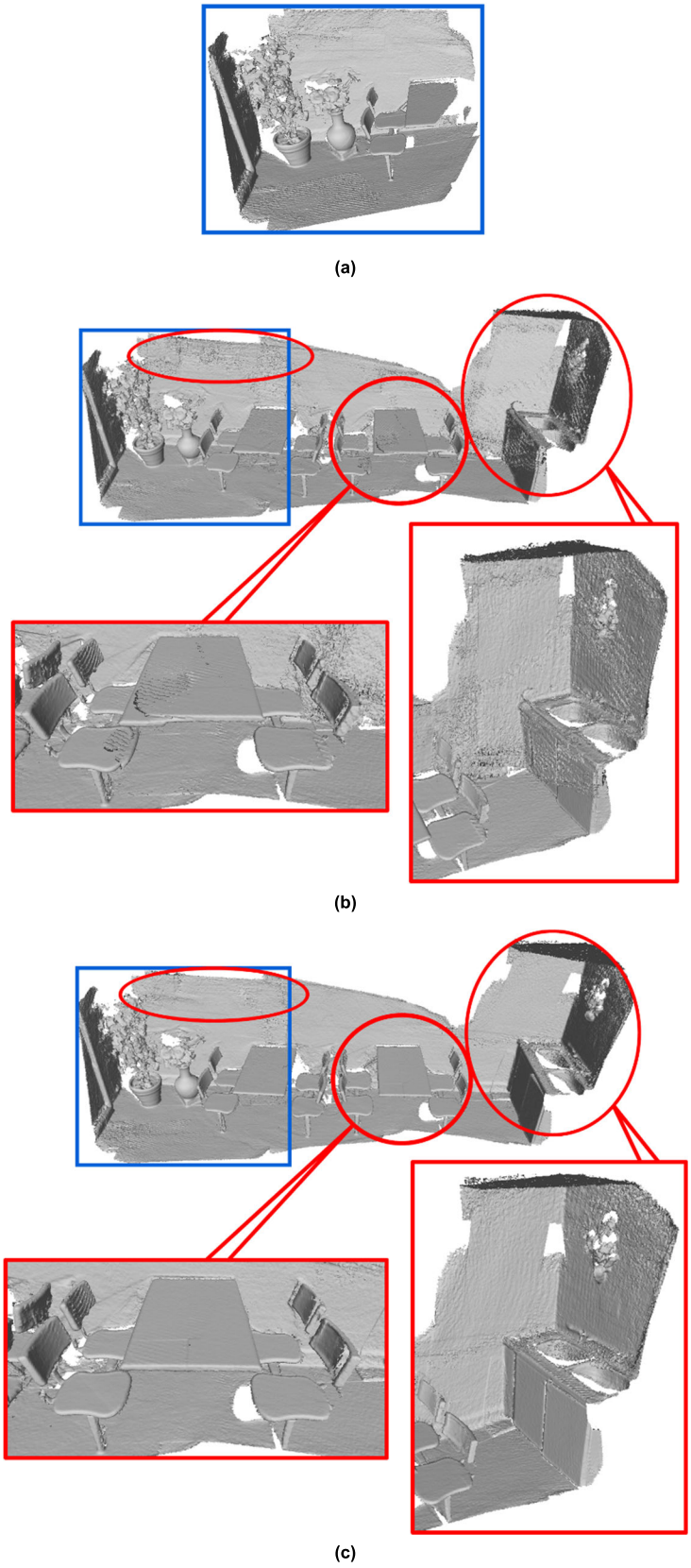


**FIGURE 11.** 3D geometry reconstruction results of: (a) LDR inputs; (b) HDR inputs; (c) normalized HDR inputs on low light cafeteria dataset.

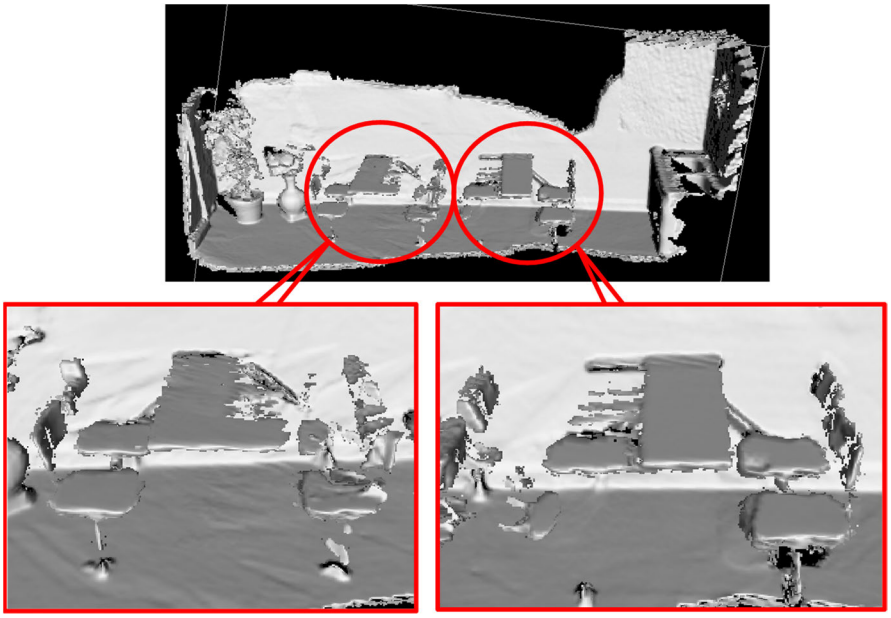

**FIGURE 12.** 3D geometry reconstruction result of HDRFusion [14] on low light cafeteria dataset.

RGB-D frame resolution is 640 × 480; resolution of TSDF volume is 512×512×512; size of TSDF volume is 10 (meter); initial camera position with respect to TSDF volume is [3.0, 3.0, 3.0] (meter). Noted that since HDRFusion [14] is heavily based on camera calibration process to generate HDR images, we could only carry out the experiment with our own dataset (low light cafeteria) but not TUM datasets due to the lack of their camera calibration information.

Fig. 12 shows that for the low light cafeteria dataset, due to the constraint of GPU memory, smaller TSDF volume resolution leads to coarser reconstruction. Additionally, based on the observation that 3D geometries of the tables and the chairs are mostly incomplete, we can conclude that the reconstruction quality of our proposed method is better.

## VI. CONCLUSION

This paper has presented a robust normalized HDR-based 3D reconstruction pipeline to reconstruct challenging low light scenes scanned with a consumer RGB-D depth camera. Different from related methods which rely on pre-calibrated CRF function to generate HDR images from LDR image, our adopted deep learning-based generation method is not restricted by specific calibrated camera and thus has better applicability. We have evaluated the proposed method on challenging low light TUM RGB-D dataset and our dataset. Experimental results show that the proposed normalized HDR-based 3D reconstruction method performs better than ORB-SLAM2 with LDR inputs and HDR-SLAM with HDR inputs in terms of both camera tracking accuracy and 3D geometry reconstruction quality. We also demonstrated that compared to ORB-SLAM2 which is prone to lose tracking and reconstruct defect models, the proposed method can successfully track all the frames and is robust to low light environment, low texture, and fast motion. The proposed method has demonstrated the following two concepts: besides dense SLAM, it is also feasible and beneficial to use HDR images as the input of feature-based SLAM since salient features can be detected in HDR images; second, the deep learning based HDR image generation framework can be adopted in various system to replace complex camera calibration process.

In the future, since we have demonstrated the effectiveness of incorporating the proposed HDR-SLAM into 3D reconstruction system, we are going to add the low-light detection mechanism to build a more flexible system with minimum increase on computation complexity. In addition, we will extend this work by incorporating normalized radiance map and learning-based deep fusion method into the surface reconstruction stage.

## REFERENCES

[1] G. Tiwari and P. Rani, "A review on high-dynamic-range imaging with its technique," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 9, pp. 93–100, Sep. 2015.
[2] A. El Gamal, "High dynamic range image sensors," in *Proc. Tutorial Int. Solid-State Circuits Conf.*, vol. 290, 2002.
[3] G. Wan, X. Li, G. Agranov, M. Levoy, and M. Horowitz, "CMOS image sensors with multi-bucket pixels for computational photography," *IEEE J. Solid-State Circuits*, vol. 47, no. 4, pp. 1031–1042, Apr. 2012.
[4] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. ACM SIGGRAPH Classes (SIGGRAPH)*, 2008, p. 31.
[5] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," *Comput. Graph. Forum*, vol. 28, no. 1, pp. 161–171, 2009.
[6] M. A. Robertson, S. Borman, and R. L. Stevenson, "Dynamic range improvement through multiple exposures," in *Proc. Int. Conf. Image Process.*, vol. 3, 2019, pp. 159–163.
[7] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2519–2532, May 2017.
[8] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–10, Nov. 2013.

[9] J. Im, J. Jeon, M. Hayes, and J. Paik, "Single image-based ghost-free high dynamic range imaging using local histogram stretching and spatially-adaptive denoising," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1478–1484, Nov. 2011.

[10] A. K. Johnson and C. V. Jiji, "Single shot high dynamic range imaging using histogram separation and exposure fusion," in *Proc. 5th Nat. Conf. Comput. Vis., Pattern Recognit., Image Process. Graph. (NCVPRIPG)*, Dec. 2015, pp. 1–4.

[11] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–15, Nov. 2017.

[12] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[13] M. Meilland, C. Barat, and A. Comport, "3D high dynamic range dense visual SLAM and its application to real-time object re-lighting," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2013, pp. 143–152.

[14] S. Li, A. Handa, Y. Zhang, and A. Calway, "HDRFusion: HDR SLAM using a low-cost auto-exposure RGB-D sensor," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 314–322.

[15] C.-H. Yeh, M.-H. Lin, and W.-C. Lu, "3D reconstruction using HDR-based SLAM," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1976–1980.

[16] C. Barat and A. I. Comport, "Active high dynamic range mapping for dense visual SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 6514–6519.

[17] S. V. Alexandrov, J. Prankl, M. Zillich, and M. Vincze, "High dynamic range SLAM with map-aware exposure time control," in *Proc. Int. Conf. 3D Vis. (DV)*, Oct. 2017, pp. 48–56.

[18] T. M. Pinho, J. P. Coelho, J. Oliveira, and J. Boaventura-Cunha, "Comparative analysis between LDR and HDR images for automatic fruit recognition and counting," *J. Sensors*, vol. 2017, pp. 1–12, Jan. 2017.

[19] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. W. H. Lau, "Image correction via deep reciprocating HDR transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1798–1807.

[20] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

[21] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald, "Kintinuous: Spatially extended KinectFusion," in *Proc. RSS Workshop RGB-D, Adv. Reasoning Depth Cameras*, Jul. 2012.

[22] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1697–1716, Sep. 2016.

[23] Q.-Y. Zhou and V. Koltun, "Dense scene reconstruction with points of interest," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–8, Jul. 2013.

[24] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5556–5565.

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

[26] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard, "Real-time 3D visual SLAM with a hand-held RGB-D camera," in *Proc. RGB-D Workshop 3D Perception Robot. Eur. Robot. Forum*, vol. 180, Apr. 2011, pp. 1–15.

[27] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[28] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 4, p. 1, Jul. 2017.

[29] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1691–1696.

[30] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad, "An overview to visual odometry and visual SLAM: Applications to mobile robotics," *Intell. Ind. Syst.*, vol. 1, no. 4, pp. 289–311, Dec. 2015.

[31] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, Aug. 1987.

[32] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! Large-scale texturing of 3D reconstructions," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 836–850.

[33] Q.-Y. Zhou and V. Koltun, "Color map optimization for 3D reconstruction with consumer depth cameras," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.

[34] C. Liu, W. T. Freeman, R. Szeliski, and S. Bing Kang, "Noise estimation from a single image," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 901–908.

[35] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, "ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 37–49, May 2018.

[36] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[37] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.

[38] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.

[39] A. Jung. *Imgaug*. Aug. 2017. [Online]. Available: https://github.com/aleju/img

[40] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 4, pp. 629–642, 1987.

[41] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: A survey from 2010 to 2016," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, p. 16, Dec. 2017.

**CHIA-HUNG YEH** (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan, in 1997 and 2002, respectively. He was an Assistant Professor from 2007 to 2010, an Associate Professor from 2010 to 2013, and a Professor from 2013 to 2017 with the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. He is currently a Distinguished Professor with National Taiwan Normal University, Taipei, Taiwan. He has coauthored more than 250 technical international conferences and journal articles and held 47 patents in the USA, Taiwan, and China. His research interests include multimedia, video communication, three-dimensional reconstruction, video coding, image/video processing, and big data. He is an Associate Editor of the *Journal of Visual Communication and Image Representation*, *EURASIP Journal on Advances in Signal Processing*, and *APSIPA Transactions on Signal and Information Processing*. He has been on the Best Paper Award Committee of JVCI and APSIPA. He was a recipient of the 2007 Young Researcher Award of NSYSU, the 2011 Distinguished Young Engineer Award from the Chinese Institute of Electrical Engineering, the 2013 Distinguished Young Researcher Award of NSYSU, the 2013 IEEE MMSP Top 10% Paper Award, the 2014 IEEE GCCE Outstanding Poster Award, the 2015 APSIPA Distinguished Lecture, the 2016 NARLabs Technical Achievement Award: Superior Achievement Award, the 2017 IEEE SPS Tainan Section Chair, the 2017 Distinguished Professor Award of NTNU, and the IEEE Outstanding Technical Achievement Award (IEEE Tainan Section).

**MIN-HUI LIN** received the B.S. degree from the Department of Electrical Engineering, National Kaohsiung University, Taiwan, in 2016. She is currently pursuing the Ph.D. degree in electrical engineering with National Sun Yat-sen University, Taiwan. Her research interests include in the areas of deep learning for computer vision, 3D reconstruction, and multimedia applications.