LSFB: A Low-cost and Scalable Framework for Building Large-Scale Localization Benchmark

Haomin Liu * [1], Mingxuan Jiang[1], Zhuang Zhang[1], Xiaopeng Huang[1], Linsheng Zhao[1], Meng Hang[1], Youji Feng[1], Hujun Bao[2] and Guofeng Zhang * [2]

[1] *SenseTime Research*
[2] *State Key Lab of CAD&CG, Zhejiang University*

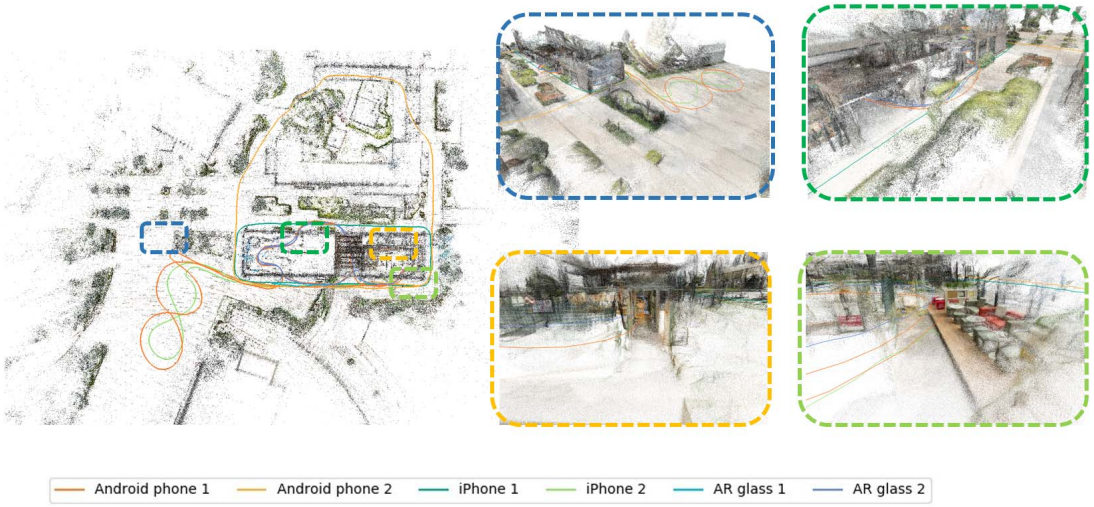

Figure 1: The reconstructed HD map overlaid with recovered trajectories of multiple AR devices moving around indoors and outdors. The top view are shown on the left, and the details are shown on the right.

ABSTRACT

With the rapid development of mobile sensor, network infrastructure and cloud computing, the scale of AR application scenario is expanding from small or medium scale to large-scale environments. Localization in the large-scale environment is a critical demand for the AR applications. Most of the commonly used localization techniques require quite a number of data with groundtruth localization for algorithm benchmarking or model training. The existed groundtruth collection methods can only be used in the outdoors, or require quite expensive equipments or special deployments in the environment, thus are not scalable to large-scale environments or to massively produce a large amount of groundtruth data. In this work, we propose LSFB, a novel low-cost and scalable framework to build localization benchmark in large-scale environments with groundtruth poses. The key is to build an accurate HD map of the environment. For each visual-inertial sequence captured in it, the groundtruth poses are obtained by joint optimization taking both the HD map and visual-inertial constraints. The experiments demonstrate the obtained groundtruth poses are accurate enough for AR applications. We use the proposed method to collect a dataset of both mobile phones and AR glass exploring in large-scale environments, and will release the dataset as a new localization benchmark for AR.

*e-mails: liuhaomin@sensetime.com, zhangguofeng@cad.zju.edu.cn.

Index Terms: Human-centered computing—Human Computer Interaction—Interaction Paradigms—Mixed / Augmented Reality; Computing Methodologies—Artificial Intelligence—Computer Vision—Tracking

1 INTRODUCTION

Localization is the core to augmented reality (AR). One of the most popular techniques is visual-inertial SLAM (VI-SLAM), that uses the complementary visual and inertial measurements to localize the mobile device in the unknown environment [27, 30]. The state-of-the-art VI-SLAM has achieved great accuracy and robustness in small and medium scale environments [7], and has been integrated into many AR products like Apple ARKit, Google ARCore and Microsoft Holenlens etc. With the rapid development of mobile sensors, network infrastructure and cloud computing, the scale of AR application scenarios is also expanding, from medium-scale rooms to city-scale environments. This trend has received unprecedented attention of both academics and industry, but also posed unprecedented challenges on the accuracy and robustness [3]. To meet the challenges, it is important to have an appropriate benchmark containing a large amount of data with groundtruth localization in the large-scale environments. Existed groundtruth collection methods either require GPS that is not available indoors [11], or require quite expensive equipments [2, 16] or special deployments [5, 28] in the environment, not scalable to large-scale environments or to massively produce a large amount of groundtruth data.

In this work, we propose **LSFB**, a novel **L**ow-cost and **S**calable **F**ramework to build localization **B**enchmark in large-scale environ-

978-1-7281-7675-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR-Adjunct51615.2020.00065

Table 1: Comparison of existed groundtruth collection methods

	groundtruth solution	equipment cost	operation cost	environment scale	indoor / outdoor	suitable for AR
KITTI [11]	GPS + IMU	medium	medium	large	no / yes	no
Rawseeds [5]	GPS + fiducial marker	medium	high	medium	yes / yes	no
NCLT [4]	GPS + IMU + Lidar	medium	medium	large	yes / yes	no
Complex Urban [15]	GPS + IMU + Lidar	medium	medium	large	yes / yes	no
PennCOSYVIO [28]	fiducial marker	low	high	medium	yes / yes	yes
ADVIO [6]	position fixes + IMU	low	high	medium	yes / yes	yes
EuRoC [2]	VICON	high	high	small	yes / no	yes
ZJU-SenseTime [16]	VICON	high	high	small	yes / no	yes
Aachen Day-Night [32]	SfM + manual annotation	low	high	large	yes / yes	no
CMU Seasons [32]	SfM + manual annotation	low	high	large	yes / yes	no
LSFB	HD Map + VI-constraints	low	low	large	yes / yes	yes

ments with groundtruth poses. As shown in Fig. 1, our framework is based on building an accurate HD map of the environment. For each visual-inertial sequence captured in the environment, the groundtruth poses is obtained by joint optimization taking both the HD map and visual-inertial constraints. Compared to the existed groundtruth collection methods, our solution is the only one with a combination of attributes including:

- low equipment cost
- low operation cost
- scalable to large environment
- suitable for both indoor and outdoor
- suitable for AR

Detailed comparison is listed in Table 1, and explained in Sect. 2.1.

2 Related Work

In this section, we first review and compare the existed groundtruth collection methods, then review the localization techniques this work focuses on.

2.1 Groundtruth Solution

The very popular KITTI [11] uses a high precision GPS/IMU to produce 6 DOF groundtruth poses in large-scale outdoor environments. To extend to indoor area where GPS is not available, Rawseeds [5], NCLT [4] and Complex Urban [15] propose to use fiducial marker or Lidar scan matching as additional pose constraint in the indoor area. In Rawseeds, the fiducial markers are attached to the moving robot, observed by an external network of pre-calibrated cameras and laser scanner. These external cameras and laser scanners increase the equipment and operation cost which limits the scalability. Opposite to this outside-in solution, both NCLT [4] and Complex Urban [15] use the inside-out measurement of Lidar scan matching, alleviating the cost and scalability limitation. However, they require high precision Lidar on the mobile device, which is infeasible for AR devices. In addition, all the above mentioned methods collect data by a wheeled ground vehicle/robot, whose motion style is quite different from to the handheld/wearable AR devices. By contrast, PennCOSYVIO [28] and ADVIO [6] use handheld VI-sensor as the localized device, thus are suitable for AR and the equipment cost is low. But they respectively require manually measured fiducial markers placed in the environment [28], and manually labeled position to fixes on the floor plan [6], to provide the groundtruth position constraints, thus the operation cost is high and the scalability is limited. EuRoC [2] and ZJU-SenseTime [16] use a VICON motion capture system, which is comprised of VICON trackers deployed in the environment and VICON markers attached to the localized device. Both equipment and deployment cost of VICON trackers are prohibitively high ($\sim$30K dollar for a 20 m^2 room) to scale to a large environments or outdoors. Aachen Day-Night and CMU Seasons proposed in [32] are two commonly used datasets for larege-scale visual localization. Groundtruth is obtained by the Structure-from-Motion (SfM) technique [34]. Considering day-night and across seasons challenges, manually labeled correspondences are required, which increases the operation cost, and the images are not sequential thus not suitable for AR.

Compared to the existed method, LSFB proposed in this work does not require any special deployment in the environments or special equipment on the localized device except a VI-sensor with synchronized camera and IMU, which is commonly available on the current AR device (e. g. mobile phones and AR glasses), thus requires low equipment and operation cost, and suitable for both indoor and outdoor environment for the AR application. The only requirement is that we assume the HD map of the environment has been accurately reconstructed. The groundtruth trajectory for each visual-inertial sequence captured in the environment are obtained by joint optimization taking both the global localization constraint from the HD map and the temporal motion constraint from visual-inertial measurements. Thanks to the recently maturing large-scale 3D reconstruction techniques [13,41] and the efficient optimization techniques [17, 22], our method is scalable to large-scale environments with a large amount of data.

2.2 Localization

Simultaneous Localization and Mapping (SLAM) is probably the most commonly used localization technique for AR. PTAM [19] is the first visual SLAM (V-SLAM) system that is able to track the camera motion in real-time on mobile devices. Bundle Adjustment (BA) [38] that jointly optimizes keyframe poses and 3D points is performed in a background thread to ensure accuracy without jeopardizing real-time performance. This framework is used by most of the recent V-SLAM systems [23, 26, 37]. Visual-inertial SLAM (VI-SLAM) combines complementary visual and inertial measurements to achieve better accuracy and robustness than V-SLAM, and becomes a standard technique for AR applications. The most successful systems include MSCKF [25], OKVIS [21], VINS [30] etc. Jeffrey Delmerico and Davide Scaramuzza give a detailed evaluation of the state-of-the-art VI-SLAM systems [7].

Visual localization is a promising technique to obtain the global localization in HD map. Traditional methods [14, 31, 33] mainly aim to improve the efficiency and accuracy of the 2D-3D correspondence search. They either use image retrieval [36] to find the correspondence in a coarse-to-fine manner, or employ quantization method [29] to perform direct feature matching. Recent approaches resort to learning-based features (e. g. SuperPoint [8], D2-Net [9], etc) to improve the robustness against view point changes, motion blurs, and illumination variations, etc. Localization can also be estimated in an end-to-end manner (e. g. pose-net [18], map-net [1] etc) thanks to the strong capacity of deep neural networks. The main limitation is the generalization issue. The poses would be terribly wrong in some viewpoints that are not covered by the training set. A large amount of training data with groundtruth localization

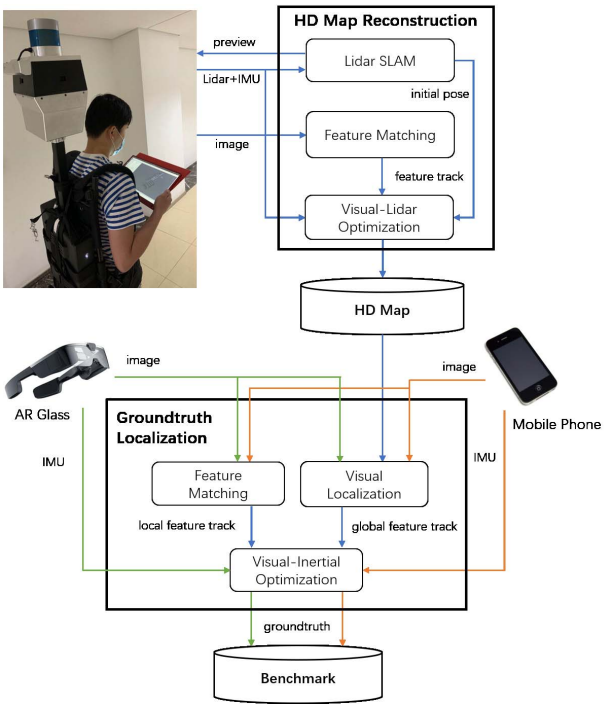

Figure 2: The framework of LSFB

is required to make these methods work.

3 METHOD

The framework of the proposed method is illustrated in Fig. 2. We developed a backpack 3D scanner comprised of a Velodyne Ultra Puck VLP-3DC 3D Lidar and a customized Mynt VI-sensor to capture the Lidar point cloud and 360° images of the environment, from which the HD map of the environment is reconstructed (Sect. 3.1). The HD map is used by all the localized devices with VI-sensor (e. g. mobile phone and AR glass) moving in the same environment. Each visual-inertial sequence is localized in the HD map, within which each frame associated with a groundtruth pose (Sect. 3.2).

3.1 HD Map Reconstruction

As shown in the upperleft picture of Fig. 2, a 3D Lidar is placed on top of a fixed structural part containing an array of 4 cameras capturing the 360° images of the environment and an IMU. The camera specification is listed in Table 2. All sensors are hardware synchronized and carefully calibrated.

To guarantee the completeness and accuracy of the final reconstruction, we use the Lidar and IMU measurements to run 3D Lidar SLAM that performs dense mapping in real-time. The reconstruction results is displayed in the handheld screen, guiding the operator to perform a complete and algorithm-friendly scanning. We select Google Cartographer [13] for its robustness, convenience and the ability to close large loops in real-time. Cartographer uses IMU measurements to project Lidar points to the latest Lidar frame at time i (called undistortion) and estimates the optimal Lidar pose ${}^{W}\mathbf{T}_{L_i}$ that best aligns the undistorted point cloud in the Lidar coordinate L_i to the global map coordinate W. The stream of 360° images, the undistorted Lidar point clouds and the Lidar poses are stored for the final offline reconstruction. The goal is to reconstruct a set of 3D points with the image feature descriptors for the visual localization. To this goal, SIFT features [24] are extracted and matched from 360° images. If the camera poses are known, the 3D position of each feature point can be obtained by triangulation [12]. The

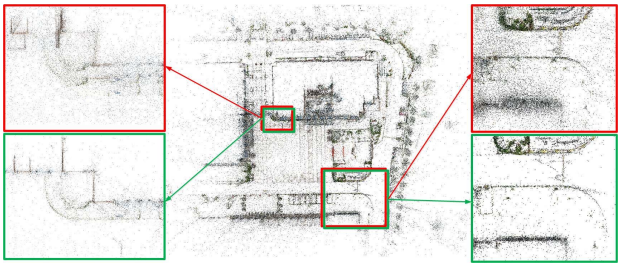

Figure 3: Top view of feature map, red box: init pose using, green box: visual-Lidar optimized.

camera pose ${}^{W}\mathbf{T}_{C_i}$ for each image i could be converted from the Lidar pose by ${}^{W}\mathbf{T}_{C_i} = {}^{W}\mathbf{T}_{L_i}\,{}^{L}\mathbf{T}_{C}$, where ${}^{L}\mathbf{T}_{C}$ is the pre-calibrated extrinsic between Lidar and camera. However, ${}^{W}\mathbf{T}_{L_i}$ from the online Cartographer is inevitably suboptimal. Directly use this pose would result in a very noisy HD map. Existed method for Lidar-based reconstruction [39] mainly focus on improving the consistency between individual Lidar point clouds, rather than improving the consistency between visual-Lidar measurements to obtain an optimal visual map. We propose a novel method taking both the visual and Lidar constraints to obtain this goal.

Given the undistorted Lidar point clouds, we first use LOAM [40] to extract plane and edge features and obtain a set of plane/edge correspondences, denoted as $\mathbb{L}_p = \{L_p = ({}^{n}\mathbf{p}, {}^{m}\mathbf{q}, {}^{m}\mathbf{n})\}$ and $\mathbb{L}_e = \{L_e = ({}^{n}\mathbf{p}, {}^{m}\mathbf{q}, {}^{m}\mathbf{l})\}$ respectively, where ${}^{n}\mathbf{p}$ is the feature point in the source point cloud n, ${}^{m}\mathbf{q}$ the matched point in the target point could m with the normal vector ${}^{m}\mathbf{n}$ for plane feature and the direction vector ${}^{m}\mathbf{l}$ for edge feature. The Lidar constraint $\mathbb{L}_p$ and $\mathbb{L}_e$ are combined with the set of image feature correspondences to jointly optimize the camera poses ${}^{W}\mathbf{T}_i$ and 3D point $\mathbf{X}_j$ of the visual map

$$\begin{aligned}\underset{{}^{W}\mathbf{T}_i, \mathbf{X}_j}{\arg\min} & \sum_{L_p \in \mathbb{L}_p} \| {}^{m}\mathbf{n}^{T} \cdot ({}^{L_m}\mathbf{T}_{L_n} \circ {}^{n}\mathbf{p} - {}^{m}\mathbf{q}) \|^2_{\Sigma_p} \\ + & \sum_{L_e \in \mathbb{L}_e} \| {}^{m}\mathbf{l} \times ({}^{L_m}\mathbf{T}_{L_n} \circ {}^{n}\mathbf{p} - {}^{m}\mathbf{q}) \|^2_{\Sigma_e} + \sum_{(\mathbf{X}_j, \mathbf{x}_{ij})} \| f({}^{W}\mathbf{T}_i, \mathbf{X}_j) - \mathbf{x}_{ij} \|^2_{\Sigma_v}\end{aligned} \tag{1}$$

where the 3 residuals are Lidar point-to-plane / point-to-edge distance error and image feature reprojection error. $f(\mathbf{T}, \mathbf{X})$ projects a 3D point $\mathbf{X}$ to the image plane by camera pose $\mathbf{T}$. $\mathbf{x}_{ij}$ is the 2D feature point of the $\mathbf{X}_j$ in image i. Σ_p, Σ_e and Σ_v are the covariance matrixes for plane, edge and image feature constraint, which is set to be 0.2 m^2, 0.5 m^2 and 2 $pixel^2$ respectively. (1) is solved iteratively using the LM algorithm. The reconstructed visual map is visualized in Fig. 3. It can be observed that directly using the initial pose results in very noisy visual map. With the proposed visual-Lidar opitmization, the visual map is more accurate.

3.2 Groundtruth Localization

For each input frame, visual localization is performed to find a set of feature correspondences between the 2D feature point in the input image and the 3D feature point in the HD map. We call them global feature correspondences/tracks. The camera pose can be estimated from these 2D-3D correspondences by PnP [20]. However, the result of PnP highly relies on the number of feature correspondences and their distribution in the image. It is inevitable that there are viewpoints that are not covered by the HD map, leading to localization failure in the uncovered area. In addition, the pose obtained by PnP frequently jitters due to lack of temporal constraints, thus is not suitable for AR with strict requirements for smoothness. We propose to use the visual-inertial measurements in the localized sequence as the temporal constraint. To this purpose, we extract and match features for the input image sequence. We call them lo-

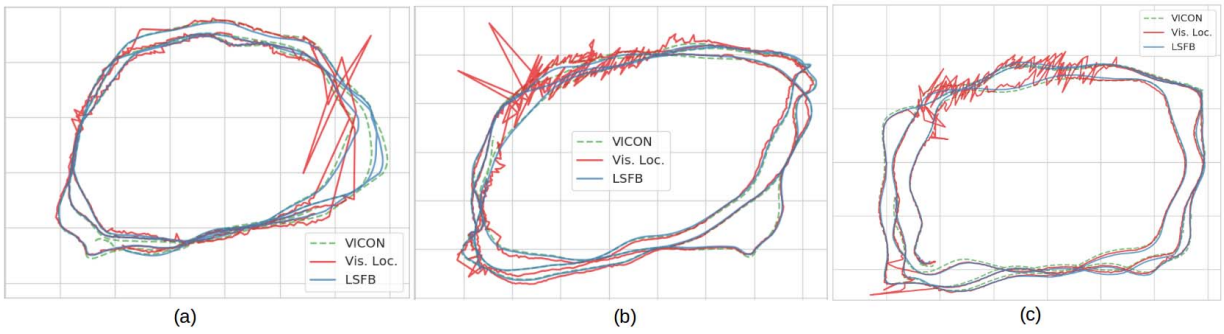

Figure 4: The trajectories in the VICON room for (a) iPhone, (b) Android phone and (c) AR glass.

cal feature correspondences/tracks. We select both KLT [35] and SIFT [24] feature for their complementary property, i. e. better robustness of KLT, and better accuracy of SIFT and with the ability to close loops. IMU measurements are also used to provide temporally constraint between consecutive poses by IMU integration [10]. The problem is formulated as a joint optimization to find the optimal set of camera motion parameters for each frame and 3D position for each local feature track, i. e. the Bundle Adjustment (BA) [38], satisfying both the global localization constraint from HD map and the temporal motion constraint from visual-inertial measurement.

$$\underset{\mathbf{C}_i,\mathbf{X}_j}{\arg\min} \sum_i \left(\sum_j \|f({}^W\mathbf{T}_i, \mathbf{X}_j^*) - \mathbf{x}_{ij}^*\|_2^2 + \sum_j \|f({}^W\mathbf{T}_i, \mathbf{X}_j) - \mathbf{x}_{ij}\|_2^2 \right) + \sum_i \|h(\mathbf{C}_i, \mathbf{C}_{i+1})\|_2^2 \tag{2}$$

where $\mathbf{C}_i = ({}^W\mathbf{T}_i, {}^W\mathbf{v}_i, \mathbf{b}_a, \mathbf{b}_g)$ are camera variables for image i, comprised of the camera pose ${}^W\mathbf{T}_i$, the velocity ${}^W\mathbf{v}_i$ and IMU bias $\mathbf{b}_a$ and $\mathbf{b}_g$ for acceleration and gyroscope measurement respectively. $\{(\mathbf{X}_j^*, \mathbf{x}_{ij}^*)\}/\{(\mathbf{X}_j, \mathbf{x}_{ij})\}$ are the set of global/local feature correspondences found in image i, with $\mathbf{X}_j^*$ the fixed 3D point in the HD map and $\mathbf{X}_j$ the 3D point of local feature BA aims to optimize. $h(\cdot)$ is the IMU cost function evaluating the difference between the relative camera motion and the pre-integration result calculated from IMU measurements for each consecutive camera pair [10].

BA is essentially a non-linear optimizer, which is sensitive to initial value. It is best to perform BA after each frame is added to (2) to give the optimal initial value to the next run, but it is obviously computational infeasible for traditional BA solver to support long-time exploration in large-scale environments. We leverage the keyframe strategy that is widely used in current SLAM systems to reduce the computation complexity [19, 23, 26]. For each incoming frame i, if it observes less than 35 existed local features, it will be selected as a new keyframe, and triggers the global BA performed on all keyframes. Otherwise we record the transformation ${}^k\mathbf{T}_i$ from i to its nearest keyframe k, which is the one shares most common features with i. After all frames are processed, we recover the camera pose for each non-keyframe i by ${}^W\mathbf{T}_i = {}^W\mathbf{T}_k\,{}^k\mathbf{T}_i$, and perform a final global BA (2) on all keyframe and non-keyframes.

Even with the keyframe strategy, the bottleneck still exists since the complexity of BA grows quadratically with the number of keyframes. It will quickly become computational infeasible to perform global BA for each incoming keyframe. We use the recently developed ICE-BA [22] to break this limitation. ICE-BA is a BA solver specialized for the VI-SLAM problem in which the incoming visual-inertial measurements arrive sequentially. ICE-BA explicitly leverages this fact and re-uses the intermediate results of previous optimization to avoid redundant new computation, to achieve the exact solution with significantly higher efficiency.

Table 2: Specification of cameras

	Backpack 3D scanner	AR device		
		iPhone	AN. phone	AR glass
#Cam.	4	1	1	2
Res.	1280×800	640×480	640×480	640×400
FoV	106°×65°	68°×54°	66°×52°	97°×70°
Freq.	20 Hz	30 Hz	30 Hz	30 Hz

Table 3: The RMSE comparison

	LSFB	Vis. Loc.	LSFB w/ Pct. of Vis. Loc.			
			60%	40%	20%	0%
iPhone	0.037	0.068(65.08%)	0.039	0.036	0.055	0.124
AN.phone	0.046	0.112(97.55%)	0.044	0.044	0.025	0.104
AR glass	0.044	0.088(98.43%)	0.037	0.033	0.039	0.133

4 Evaluation

In this section, we first conduct experiments to analyze the accuracy of the proposed method quantitatively, then give the qualitative result of the collected groundtruth dataset which will be released as a new localization benchmark for AR. We use the backpack 3D scanner to reconstruct the HD map of the environments, and use 3 AR devices to collect the visual-inertial data in the environment. The 3 AR devices are: an iPhone XS Max, an Android phone Huawei Mate 30 Pro, and a Shadow Creator AR glass Action One Pro. The camera specifications for the backpack 3D scanner and the AR devices are listed in Table 2.

4.1 Quantitative Result

To the best of our knowledge, there is no other way to evaluate the localization accuracy in large-scale environments for the mobile phones and AR glass. We can only conduct quantitative evaluation in a small room (∼20 m^2), using VICON to obtain the localization with millimeter accuracy as the baseline to evaluate our method.

For each of the 3 AR devices, we walked around the room 3 times at the speed of ∼1 m/s. The camera poses estimated by LSFB are compared to that measured by VICON. The translational RMSE (Root Mean Square Error) is calculated for each sequence, and listed in Table 3. We can see that LSFB achieves centimeter-level accuracy for all devices. We also evaluate the camera poses obtained by visual localization. The RMSE are listed in the next column, and the trajectories are shown in Fig. 4. As expected, the poses from visual localization is erroneous and frequently jitters, which is not suitable for AR which has strict requirements for the localization smoothness. We also count the success rate of visual localization in the bracket, which is defined as the percentage of frames that have matched more than 30 features in the HD map. The success rate is high for Android phone and AR glass, but lower for iPhone because the focus is locked, resulting in severe image blur. By the proposed LSFB, all the poses can be accurately recov-

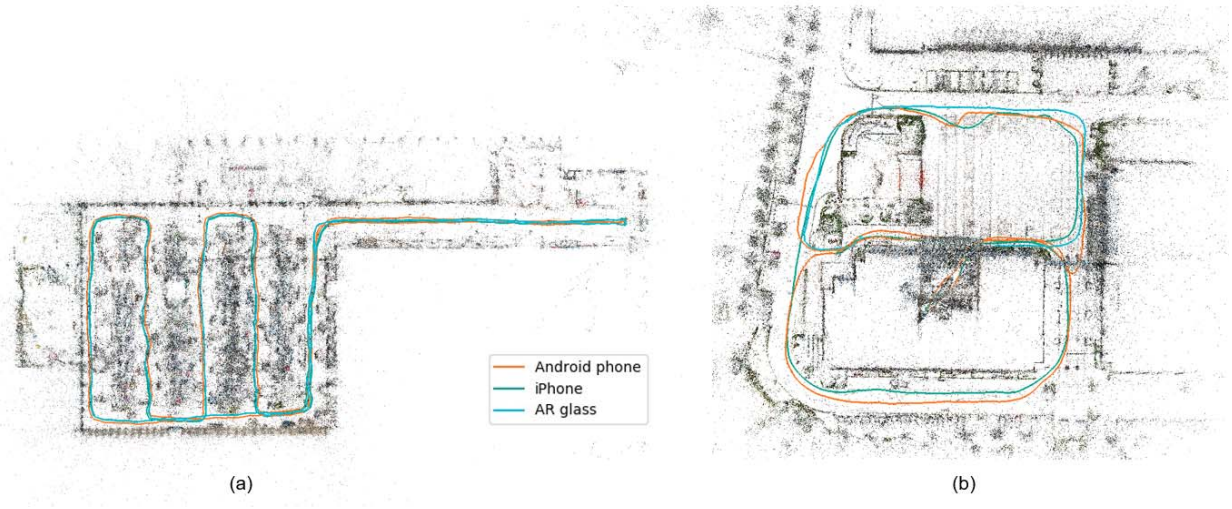

Figure 5: The reconstructed HD map overlaid with recovered trajectories in (a) "indoor" and (b) "outdoor".

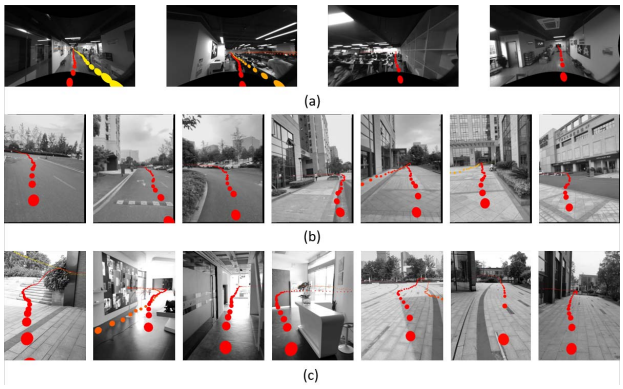

Figure 6: The snapshots of 3 benchmark sequences in 3 different environments, (a) AR glass in "indoor", (b) iPhone in "outdoor" and (c) Android phone "indoor-outdoor". Future trajectory represented as a sequence of virtual spheres is overlaid in the images to demonstrate the AR effect.

ered.

It is easy to fully cover the small room by the 360°images, but not the case for large-scale environments. We conduct an experiment to investigate the effect of visual localization failure. We randomly select a portion of frames and remove all global feature correspondences on them, resulting 60%, 40% and 20% of visual localization. We perform LSFB using the remaining global feature correspondences, and list the RMSE in the last 3 columns in Table 3. We can see that the accuracy is not sensitive to visual localization failure, sometimes even better after removing the erroneous global feature correspondences. In the last column, we also evaluate the result without HD map (0%), which is actually the traditional VI-SLAM. The RMSE is several times larger than LSFB, verifying the effectiveness of incorporating the HD map.

4.2 Qualitative Result

We collect groundtruth data in 3 environments as the LSFB benchmark for AR. The 3 environments include an indoor office $\sim$500 m^2, an outdoor environment around a building and a square $\sim$10,000 m^2 in total, and an environment containing both indoor office building and outdoor garden $\sim$10,000 m^2 in total. We call them "indoor", "outdoor" and "indoor-outdoor" respectively. We walked around the environments to collect a set of visual-inertial data using mobile phones and AR glass. We use the proposed LSFB to reconstruct the HD map and localize each sequence in it, and overlay the result together. The results are shown in Fig. 5(a)(b), and Fig. 1 for the 3 environment respectively. The trajectories are globally consistent thanks to the global feature constraints, and temporally smooth thanks to the local visual-inertial constraints.

Since the LSFB is designed for AR, we also show the AR effect using the recovered camera poses. Specifically, for each image, we overlay the future trajectory on the image, as a sequence of spheres sampled every 5 frames in the original sequence. Snapshots are shown in Fig. 6. See the supplementary video for the whole sequence. We can see that the result are smooth and accurate enough for AR applications.

5 Conclusion

In this work, we present LSFB, a novel groundtruth collection method for localization in large-scale environments. To the best of our knowledge, it is the first groundtruth framework with low-cost and the scalability to collect a large amount of data in the large-scale environments. The experiments demonstrate the obtained groundtruth poses are accurate enough for AR applications. We use the proposed method to collect a dataset of both mobile phones and AR glass exploring in large-scale environments, and will release the dataset as a new localization benchmark for AR.

The current solution relies on the precision of visual localization. If the visual localization module returns too many false positive results, the visual-inertial optimization would diverge. Another limitation is the requirement of people to carry a heavy backpack to collect HD maps, so the size of current maps is only building-scale. For the future work, we will address the failure cases, and setup the 3D scanner to a vehicle to enable city-scale HD map reconstruction before collecting the city-scale localization data to further expand our benchmark. We hope that would truly promote the localization techniques and the realization of AR applications in the city-scale environments.

Acknowledgments

This work was partially supported by NSF of China (No. 61672457), and the Fundamental Research Funds for the Central Universities (No. 2019XZZX004-09).

References

[1] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz. Geometry-aware learning of maps for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2018.

[2] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016.

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

[4] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016.

[5] S. Ceriani, G. Fontana, A. Giusti, D. Marzorati, M. Matteucci, D. Migliore, D. Rizzi, D. G. Sorrenti, and P. Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Autonomous Robots*, 27(4):353, 2009.

[6] S. Cortés, A. Solin, E. Rahtu, and J. Kannala. Advio: An authentic dataset for visual-inertial odometry. In *European Conference on Computer Vision (ECCV)*, pp. 419–434, 2018.

[7] J. Delmerico and D. Scaramuzza. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2502–2509. IEEE, 2018.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[9] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, 2019.

[10] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. Georgia Institute of Technology, 2015.

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

[12] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[13] W. Hess, D. Kohler, H. Rapp, and D. Andor. Real-time loop closure in 2d lidar SLAM. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1271–1278. IEEE, 2016.

[14] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2599–2606. IEEE, 2009.

[15] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim. Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research*, 38(6):642–657, 2019.

[16] L. Jinyu, Y. Bangbang, C. Danpeng, W. Nan, Z. Guofeng, and B. Hujun. Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality. *Virtual Reality & Intelligent Hardware*, 1(4):386–410, 2019.

[17] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert. isam2: Incremental smoothing and mapping using the bayes tree. *The International Journal of Robotics Research*, 31(2):216–235, 2012.

[18] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE International Conference on Computer Vision*, pp. 2938–2946, 2015.

[19] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 225–234. IEEE, 2007.

[20] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2):155, 2009.

[21] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

[22] H. Liu, M. Chen, G. Zhang, H. Bao, and Y. Bao. Ice-ba: Incremental, consistent and efficient bundle adjustment for visual-inertial SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1974–1982, 2018.

[23] H. Liu, G. Zhang, and H. Bao. Robust keyframe-based monocular SLAM for augmented reality. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1–10. IEEE, 2016.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[25] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *2007 IEEE International Conference on Robotics and Automation*, pp. 3565–3572. IEEE, 2007.

[26] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[27] R. Mur-Artal and J. D. Tardós. Visual-inertial monocular SLAM with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017.

[28] B. Pfrommer, N. Sanket, K. Daniilidis, and J. Cleveland. PennCOSYVIO: A challenging visual inertial odometry benchmark. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3847–3854. IEEE, 2017.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

[30] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[31] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-based localization by active correspondence search. In *European Conference on Computer Vision*, pp. 752–765. Springer, 2012.

[32] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al. Benchmarking 6DOF outdoor visual localization in changing conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8601–8610, 2018.

[33] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In R. Bowden, J. P. Collomosse, and K. Mikolajczyk, eds., *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pp. 1–12. BMVA Press, 2012. doi: 10.5244/C.26.76

[34] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[35] J. Shi et al. Good features to track. In *1994 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600. IEEE, 1994.

[36] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pp. 1470–1477 vol.2, 2003.

[37] W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao. Robust monocular SLAM in dynamic environments. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 209–218. IEEE, 2013.

[38] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *International Workshop on Vision Algorithms*, pp. 298–372. Springer, 1999.

[39] D. Zhang, Z. Gong, Y. Chen, J. Zelek, and J. Li. SLAM-based multi-sensor backpack lidar systems in gnss-denied environments. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 8984–8987. IEEE, 2019.

[40] J. Zhang and S. Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, vol. 2, 2014.

[41] R. Zhang, S. Zhu, T. Fang, and L. Quan. Distributed very large scale bundle adjustment by global camera consensus. In *IEEE International Conference on Computer Vision*, pp. 29–38, 2017.