Laser Map Aided Visual Inertial Localization in Changing Environment

Xiaqing Ding[1], Yue Wang[1], Dongxuan Li[1], Li Tang[1], Huan Yin[1], Rong Xiong[1]

Abstract— Long-term visual localization in outdoor environment is a challenging problem, especially faced with the cross-seasonal, bi-directional tasks and changing environment. In this paper we propose a novel visual inertial localization framework that localizes against the LiDAR-built map. Based on the geometry information of the laser map, a hybrid bundle adjustment framework is proposed, which estimates the poses of the cameras with respect to the prior laser map as well as optimizes the state variables of the online visual inertial odometry system simultaneously. For more accurate cross-modal data association, the laser map is optimized using multi-session laser and visual data to extract the salient and stable subset for visual localization. To validate the efficiency of the proposed method, we collect data in south part of our campus in different seasons, along the same and opposite-direction route. In all sessions of localization data, our proposed method gives satisfactory results, and shows the superiority of the hybrid bundle adjustment and map optimization[1].

I. INTRODUCTION

Metric localization in a pre-built map is important for mobile robots to navigate in the environment autonomously. By convention, this problem is mostly addressed by the LiDAR-based map building and Monte Carlo localization. However, as the LiDAR is too expensive and heavy to be widely used, the demands for low-cost and light substitutes are risen.

In the recent years, camera, because of its low price and versatility, becomes the focus of the robotics community, showing its potential in vision-based navigation. Lots of progress has been made on visual inertial navigation system (VINS) as well as visual inertial simultaneous localization and mapping (VI-SLAM), validating the feasibility of vision-based pose estimation within one session. However, visual localization, which means localizing the camera in current session on a map that is built in previous sessions, is still challenging since the environment is changing across sessions. In long term, the environmental change involves semi-static objects, like parked cars, and appearance variation across seasons and time as shown in Fig. 1. These factors severely deteriorate the data association between different sessions, leading to unstable pose estimation. In this paper, we set to study this problem of visual localization across sessions in the changing environment.

Efforts have been paid to address this problem. One way is to combine multiple sessions of maps together for localization [1], [2]. Thus the features in the current session can find

[1]Xiaqing Ding, Yue Wang, Dongxuan Li, Huan Yin, Li Tang, Rong Xiong are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, P.R. China. Yue Wang is the corresponding author wangyue@iipc.zju.edu.cn. Rong Xiong is the co-corresponding author rxiong@zju.edu.cn.
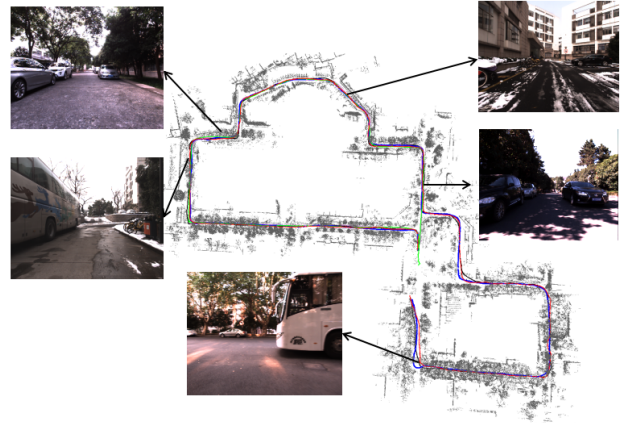[1]https://www.youtube.com/watch?v=vHAZtv2sBC4

Fig. 1: Laser map aided visual inertial localization in changing environment. The map is built using LiDAR in sunny spring, and the lines with different colors are localization trajectories in different sessions across seasons. The five pictures show some challenging circumstances during localization, which include the dynamics, semi-static objects and snow-covered landscape.

its association in the map with higher probability. The cost of this way is to build a very large map as the variations of the same place should all be saved, thus leading to heavy storage. The other way is to learn the localization based on multi-session data [3], which originates from machine learning community. The weakness of this class of methods is that the error is unstable and cannot be explained, hence still calling for further research to guide the practical application.

Recalling the laser-based solution, we can find that the core of its robustness against change in the environment owing to the geometry-based data association and estimation method, which can only be impacted by the semi-static dynamics. Inspired by this clue, the third way of visual localization [4]–[7] is promising to solve the problem, which addresses the visual localization by referring to a prior laser map. Generally there are two groups of methods to deal with this problem. One group is to synthesize images from the LiDAR built map [6], [8]. These methods inherently desire lane markings, and the computation is heavy for rendering the image and pose search. The other group is to align the local visual submap with laser map using some point cloud matching methods like iterative closest point (ICP) [9]. These methods are more general. But the sparse visual map built online can be noisy, and the dense map cannot be built efficiently without intensive computation as pointed in

[5]. In addition, the uncertainty in the local visual submap, especially on the depth direction, cannot be considered in ICP, thus introducing more error.

We model the problem of cross-modal localization into the local bundle adjustment (BA) framework to align the sparse visual submap to the LiDAR built map. In this framework, the uncertainty in both visual map and the localization between the visual and LiDAR map are optimized at the same time. Besides, we propose a saliency map extraction method based on multi-session data mining, further improving the localization performance, which also compresses the map storage for faster computation. Overall, the contributions of this paper are addressed below:

- Propose a visual inertial localization framework that addresses the localization against the laser map.
- Design a hybrid bundle adjustment which optimizes the uncertainty in local visual submap and the localization at the same time.
- Build a compressed laser map by mining the saliency from multi-session laser and visual data.
- Evaluate the proposed visual inertial localization method with experiments on multiple sessions of real-world data, which include the cross-seasonal and bidirectional circumstances.

The remainder of the paper is organized as follows: In Section II some related works about the visual localization problem are reviewed. Section III gives an overview of the whole visual inertial localization framework that includes an online visual inertial localization system and an offline laser map optimization system, which are introduced in the following Section IV and V. In Section VI we show the experimental results of the proposed visual inertial localization framework evaluated in real-world dataset. And in Section VII some conclusions are addressed.

II. RELATED WORKS

VINS provides precise and high-frequency relative pose estimation along the robot trajectory. Generally, there are two classes of methods in the area, one is to estimate the pose using a nonlinear filter, which is very efficient and light-weighted, thus is preferable in the mobile platform with limited computational resources [10], [11]. Another class of methods is based on local keyframe-based non-linear optimization, i.e. local bundle adjustment [12], [13]. The optimization based methods can achieve higher performance compared with the filter based solution, but the computational power is relatively higher. The main obstacle preventing VINS from a localization solution is that, VINS has drift in long term, since the accumulated error cannot be eliminated by itself. Therefore, in localization system, VINS is usually employed as a dead reckoning front end to support other back end.

For visual map aided localization back end [14]–[17], the maps usually contain the 3D landmarks as well as the corresponding feature-described images. While the robot is tracking with respect to the current local inertial frame, which is the origin of the VINS, the transformation between the local frame and the map frame is computed using some 2D-3D or 2D-2D matching approaches. This matching process can be achieved by either retrieving the similar images from the map dataset [18] or directly retrieving the correspondence with the whole map features [16]. To relieve the computational burden, some approaches for fast retrieving [3], [18] and map compression [15], [19] are studied. These methods enable the localization of the robot in a pre-built visual map, of which the common crucial step is the data association. For long term tasks in changing environment, the data association on general feature descriptors cannot give satisfactory performance, thus leading to the failure of the localization.

To relieve the appearance based data association, localizing the robot in a priori LiDAR map with visual sensors is considered as a useful addition [20] to compensate for the mutability of visual maps. Since visual and laser information are represented in different modalities, there are generally two ways to align vision-tracked trajectories with the laser map [7]. One way [6]–[8] is synthesizing 2D images from the 3D laser map using the intensity or depth information. The currently observed images are matched with the synthesized ones to calculate the relative transformations. To render a high quality image for accurate pose estimation, the LiDAR should be very dense. The accurate pose estimation requires high-resolution search. Thus both steps are highly computational expensive. In addition, this method inherently prefers obvious labels, such as lane markings, thus more appropriate for on-road autonomous driving. The other way [4], [21] is to reconstruct 3D points from the sequential visual images and match them with the priori 3D map to compute the relative transformation between local frame and the laser map frame. Some of these methods apply loosely coupled approaches to align the two kinds of point clouds using point cloud matching methods after visual local bundle adjustment [4]. But compared with laser point clouds, the reconstructed local visual submap is sparse and more noisy on depth direction, which cannot be modeled in ICP. In this paper, a hybrid adjustment framework is proposed to achieve the uncertainty reduction in both visual map and the localization. Different from both groups of methods, our method also employ a map compression step to improve the data association and save the storage.

III. SYSTEM OVERVIEW

Throughout this paper, the local coordinate frame is denoted as $\mathcal{L}$ and the map coordinate frame as $\mathcal{G}$. Meanwhile the current camera frame is denoted as $\mathcal{C}$, laser frame as $\mathcal{F}$ and Inertial Measurement Unit (IMU) frame as $\mathcal{B}$. We represent the pose of the robot as elements of $\mathfrak{se}(3)$, which is denoted as $\xi \in \mathbb{R}^6$. Also the operators $\mathbf{Exp}(\cdot)$ and $\mathbf{Log}(\cdot)$ as defined in [13] are also included. For each keyframe i, the corresponding linear velocity $\mathbf{v_i} \in \mathbb{R}^3$, the IMU bias terms $\mathbf{b_{ai}} \in \mathbb{R}^3$ for acceleration measurement, $\mathbf{b_{gi}}$ for gyroscope measurement together with the pose ξ_i defined in the local frame $\mathcal{L}$ of the current session, make up the state variable $\mathbf{s_i} = \{\xi_i, \mathbf{v_i}, \mathbf{b_{ai}}, \mathbf{b_{gi}}\}$.
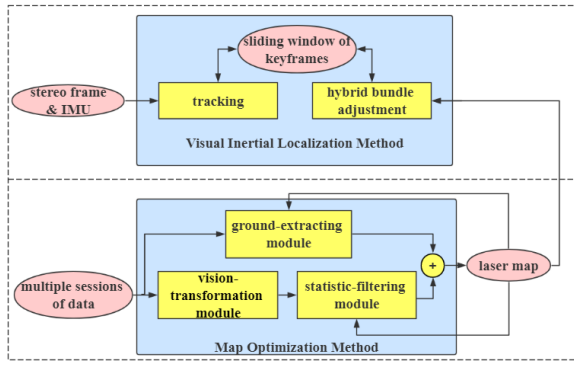
Fig. 2: The overview of the laser map aided visual inertial localization framework.

The proposed laser map aided visual inertial localization framework contains an online visual inertial localization method and an offline map optimization method as shown in Fig. 2. The online localization method tracks the camera pose $\xi_{\mathbf{i}}$ in the current session using keyframe-based visual inertial odometry using IMU preintegration. To align the trajectory to the map frame $\mathcal{G}$, the relative transformation $\xi_{\mathcal{L}}^{\mathcal{G}}$ between the local and the map frame is introduced as "anchor node" [22] which is initialized with a coarse value. When a new keyframe is created, the reconstructed visual points observed by all of the keyframes within the current sliding-window are transformed into the map frame $\mathcal{G}$ using the current estimation of $\xi_{\mathcal{L}}^{\mathcal{G}}$ to search their corresponding data association in the laser map respectively. Those matches are designed as constraints which could be added into the local BA along with the variable $\xi_{\mathcal{L}}^{\mathcal{G}}$. During optimization, the state variables in the sliding-window, local visual map points and $\xi_{\mathcal{L}}^{\mathcal{G}}$ are alternatively optimized together to achieve accurate localization of the robot in the optimized laser map.

The offline map optimization method is designed to discover the salient and stable subset within laser maps according to multiple sessions of laser and visual data. The notations of the maps are defined as follows: The visual map points represented in local frame $\mathcal{L}$ are denoted as $\mathcal{M}^{v,\mathcal{L}} = \{p_0^v, p_1^v...p_n^v\}$, where n is the number of map points and the second term in the superscript demonstrates the coordinate frame in which the variable is represented, which could be omitted if there is no ambiguity. The laser map points are all represented in the map frame $\mathcal{G}$ and are denoted as $\mathcal{M}^l = \{p_0^l, p_1^l...p_n^l\}$. The vision-transformed laser map points, which are subsets of $\mathcal{M}^l$, are denoted as $\mathcal{M}^{vl} = \{p_0^{vl}, p_1^{vl}...p_n^{vl}\}$. The aim of the map optimization method are fed with $\mathcal{M}^l$ and $\mathcal{M}^v$ to build $\mathcal{M}^{vl}$, which is expected to improve the localization accuracy by removing the ambiguous laser points. The map optimization method is a staged filter, consisting of the following three modules as shown in Fig. 2:

- **Vision-transformation module**, which removes the irrelevant laser points in each session of map $\mathcal{M}^l$ with respect to visual localization problem.

- **Statistic-filtering module**, which filters out the dynamic and semi-static points in the vision-transformed map $\mathcal{M}^{vl}$ after accumulating multiple sessions of data.
- **Ground-extracting module**, which extracts the ground points from each session of data to modify the spatial distribution of the filtered laser map $\mathcal{M}^{vl}$.

IV. VISUAL INERTIAL LOCALIZATION ON LASER MAP

Within the proposed visual inertial localization method, a keyframe-based visual inertial odometry with preintegration is applied to track the current pose w.r.t. the local frame $\mathcal{L}$. Meanwhile, a hybrid bundle adjustment is designed to align the local frame $\mathcal{L}$ to the map frame $\mathcal{G}$ to achieve the localization.

A. Visual inertial odometry

IMU provides complement measurement in some degenerated situation for vision-based odometry and provides observable pitch and roll angles. Generally it works at a much higher frequency than camera. We follow the IMU initialization and keyframe-based visual inertial tracking methods described in [23]. The preintegration between two consecutive keyframes are denoted as $[\mathbf{Log}(\Delta\mathbf{R}); \Delta\mathbf{p}; \Delta\mathbf{v}]^T$, where $\mathbf{p}$ and $\mathbf{R}$ are the translation and rotation part of ξ. The discrete evaluation of the preintegration could be derived as [13], [24]

$$
\begin{aligned}
\mathbf{R}_{i+1}^{\mathcal{LB}} =& \mathbf{R}_i^{\mathcal{LB}}\Delta\mathbf{R}_{i,i+1}\mathbf{Exp}(\mathbf{J}_{\Delta\mathbf{R}}^g\mathbf{b}_{g,i}) \\
\mathbf{v}_{i+1}^{\mathcal{B},\mathcal{L}} =& \mathbf{v}_i^{\mathcal{B},\mathcal{L}} + \mathbf{g}^{\mathcal{L}}\Delta t_{i,i+1} + \\
& \mathbf{R}_i^{\mathcal{LB}}(\Delta\mathbf{v}_{i,i+1} + \mathbf{J}_{\Delta\mathbf{v}}^g\mathbf{b}_{g,i} + \mathbf{J}_{\Delta\mathbf{v}}^a\mathbf{b}_{a,i}) \\
\mathbf{p}_{i+1}^{\mathcal{B},\mathcal{L}} =& \mathbf{p}_i^{\mathcal{B},\mathcal{L}} + \mathbf{v}_i^{\mathcal{B},\mathcal{L}}\Delta t_{i,i+1} + 0.5\mathbf{g}^{\mathcal{L}}\Delta t_{i,i+1}^2 + \\
& \mathbf{R}_i^{\mathcal{LB}}(\Delta\mathbf{p}_{i,i+1} + \mathbf{J}_{\Delta\mathbf{v}}^g\mathbf{b}_{g,i} + \mathbf{J}_{\Delta\mathbf{v}}^a\mathbf{b}_{a,i})
\end{aligned}
\tag{1}
$$

where the Jacobian $\mathbf{J}_{(\cdot)}^g$ and $\mathbf{J}_{(\cdot)}^a$ represent the first-order approximation of the effect if the variable $(\cdot)$ is changed.

Besides tracking, we keep a sliding-window of a fixed size for bundle adjustment. The cost function could be represented as

$$
\mathbf{E} = \sum_{\mathbf{s}_i \in \mathbf{S}, p_j^v \in \mathcal{M}^v} \mathbf{E}_{ba}(i,j) + \sum_{\mathbf{s}_i, \mathbf{s}_k \in \mathbf{S}} \mathbf{E}_{preint}(i,k)
\tag{2}
$$

where $\mathbf{S}$ is the set of all of the evaluated states. $\mathbf{E}_{ba}$ represents the reprojection error term between the state $\mathbf{s}_i$ and the visual point p_j^v

$$
\mathbf{E}_{ba}(i,j) = \rho((\pi(p_j^v, \xi_i) - u_{i,j})^T\mathbf{\Omega}_{i,j}^{ba}(\pi(p_j^v, \xi_i) - u_{i,j}))
\tag{3}
$$

where $\rho(\cdot)$ is the robust kernel and $\pi(\cdot, \cdot)$ is the projection function that projects p_j^v onto the image attached to pose ξ_i. $u_{i,j}$ denotes the keypoint matched with the visual point p_j^v on the image attached to pose ξ_i. $\mathbf{\Omega}_{i,j}^{ba}$ is the information matrix w.r.t. the residual between the projected point and its corresponding keypoint. $\mathbf{E}_{preint}(i,k)$ represents the
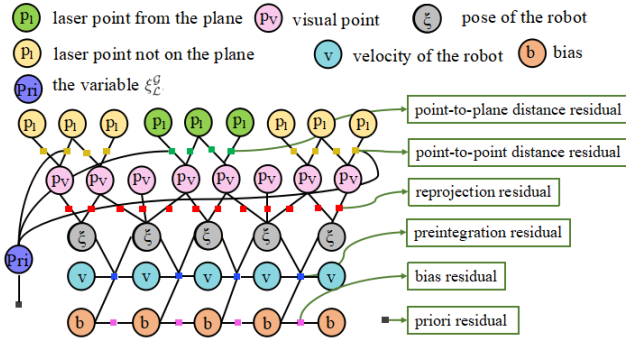
4796

Fig. 3: This graph demonstrates the proposed tightly coupled bundle adjustment. For representation some lines linked to the variable $\xi_{\mathcal{L}}^{\mathcal{G}}$ are omitted.

preintegration error term

$$
\begin{aligned}
\mathbf{E}_{preint}(i,k) &= \rho([\mathbf{e}_R^T, \mathbf{e}_p^T, \mathbf{e}_v^T]\mathbf{\Omega}_{i,k}^{pre}[\mathbf{e}_R^T, \mathbf{e}_p^T, \mathbf{e}_v^T]^T) \\
&\quad + \rho(\mathbf{e}_b^T\mathbf{\Omega}_{i,k}^{bias}\mathbf{e}_b) \\
\mathbf{e}_R &= \mathbf{Log}((\Delta\mathbf{R}_{ik}\mathbf{Exp}(\mathbf{J}_{\Delta\mathbf{R}}^g\mathbf{b}_k^g))^T)\mathbf{R}_i^{\mathcal{BL}}\mathbf{R}_k^{\mathcal{LB}} \\
\mathbf{e}_p &= \mathbf{R}_i^{\mathcal{BL}}(p_k^{\mathcal{B},\mathcal{L}} - p_i^{\mathcal{B},\mathcal{L}} - \mathbf{v}_i^{\mathcal{B},\mathcal{L}} - 0.5\mathbf{g}^{\mathcal{L}}\Delta t_{ik}^2) \\
&\quad - (\Delta\mathbf{p}_{ik} + \mathbf{J}_{\Delta p}^g\mathbf{b}_{gi} + \mathbf{J}_{\Delta p}^a\mathbf{b}_{ak}) \\
\mathbf{e}_v &= \mathbf{R}_i^{\mathcal{BL}}(\mathbf{v}_k^{\mathcal{B},\mathcal{L}} - \mathbf{v}_i^{\mathcal{B},\mathcal{L}} - \mathbf{g}^{\mathcal{L}}\Delta t_{ik}) - \\
&\quad (\Delta\mathbf{v}_{ik} + \mathbf{J}_{\Delta v}^g\mathbf{b}_{gk} + \mathbf{J}_{\Delta v}^a\mathbf{b}_{ak}) \\
\mathbf{e}_b &= [\mathbf{b}_{ai}^T, \mathbf{b}_{gi}^T]^T - [\mathbf{b}_{ak}^T, \mathbf{b}_{gk}^T]^T
\end{aligned}
\tag{4}
$$

where $\mathbf{\Omega}_{i,k}^{pre}$ is the information matrix of the preintegration and $\mathbf{\Omega}_{i,k}^{bias}$ of the bias.

B. Localization against laser map with hybrid adjustment

To align the local frame $\mathcal{L}$ to the map frame $\mathcal{G}$ where the optimized laser map exists, we introduce the relative transformation $\xi_{\mathcal{L}}^{\mathcal{G}}$ as a variable for estimation and design a hybrid bundle adjustment method which is based on the geometry information of the laser map to eliminate the drift.

At the beginning of the localization, a coarse guess of the relative transformation $\xi_{\mathcal{L}}^{\mathcal{G}}$ is given for initialization. When a new keyframe is created, all of the visual points $\mathcal{M}^v$ observed by the keyframes in the sliding-window are transformed based on the estimation of $\xi_{\mathcal{L}}^{\mathcal{G}}$ into the map frame $\mathcal{G}$. To find correspondence between two point clouds, k nearest neighbors (NN) of each visual point are searched in the optimized laser map $\mathcal{M}^{vl}$. According to the consistency among the normals of the searched laser points , we define two types of the error term corresponding to the point-to-point and point-to-plane metrics. If the normals of the searched laser points are consistent, the error term is designed as the point-to-plane distance error $\mathbf{E}_{pl}$; otherwise the error term will be related to the point-to-point error $\mathbf{E}_{pt}$.

$$
r_n(k,j) = (p_j^{vl} - \xi_{\mathcal{L}}^{\mathcal{G}}(p_k^v))^T \cdot \mathbf{n}_{p_j^{vl}}
\tag{5}
$$

$$
\mathbf{E}_{pl}(k,j) = \rho((r_n(k,j)n_{p_j^{vl}})^T\Omega_{k,j}^{pl}(r_n(k,j)n_{p_j^{vl}}))
\tag{6}
$$

$$
\mathbf{E}_{pt}(k,j) = \rho((p_j^{vl} - p_k^v)^T\Omega_{k,j}^{pt}(p_j^{vl} - p_k^v))
\tag{7}
$$

where the $\mathbf{n}_{p_j^{vl}}$ is the normal vector of p_j^{vl}.

The relative pose from local to the map frame estimated in the last step is denoted as $\bar{\xi}_{\mathcal{L}}^{\mathcal{G}}$, which is introduced as prior to constrain $\xi_{\mathcal{L}}^{\mathcal{G}}$. The corresponding error term is defined as

$$
\mathbf{E}_{prior} = \rho(\mathbf{Log}(\bar{\xi}_{\mathcal{L}}^{\mathcal{G}-1}\xi_{\mathcal{L}}^{\mathcal{G}})^T\mathbf{\Omega}_{prior}\mathbf{Log}(\bar{\xi}_{\mathcal{L}}^{\mathcal{G}-1}\xi_{\mathcal{L}}^{\mathcal{G}}))
\tag{8}
$$

If we add these error terms into the cost function of bundle adjustment (2) as shown in Fig. 3, all of the state variables, local visual map points $\mathcal{M}^v$ as well as $\xi_{\mathcal{L}}^{\mathcal{G}}$ could be optimized together, during which the uncertainties of the visual map points and the anchor point $\xi_{\mathcal{L}}^{\mathcal{G}}$ are optimized simultaneously aided with the information from the laser map. The cost function is represented as

$$
\begin{aligned}
\mathbf{E} = \sum \mathbf{E}_{ba}(i,j) &+ \sum \mathbf{E}_{preint}(i,k) \\
&+ \sum \mathbf{E}_{pl} + \sum \mathbf{E}_{pt} + \sum \mathbf{E}_{prior}
\end{aligned}
\tag{9}
$$

In our work, the Cauchy Loss is utilized as the robust kernel and this nonlinear least square optimization problem is solved based on Levenberg-Marquardt algorithm [25]. Since all of the variables are optimized together, this process could be referred as **tightly coupled bundle adjustment**

During the tightly coupled bundle adjustment, all of the information is utilized for optimization, which leads to high performance of estimation and relieves the accuracy of data association. But if the initial value is not approaching the minima, the whole localization process may result in poor estimation. On the other side, map alignment could also be achieved by applying ICP after the general bundle adjustment, which means the variables of visual inertial odometry and the anchor point $\xi_{\mathcal{L}}^{\mathcal{G}}$ will be estimated separately. We refer this method as **loosely coupled bundle adjustment**. Note that the method in [4] belongs to this class, thus is stated in the framework as a special case. Considering the estimation of bias and velocity are separated from localization stage, the loosely coupled bundle adjustment is expected to be more robust to initial value.

To combine the accuracy of the tightly coupled bundle adjustment and the robustness to initial value of the loosely coupled bundle adjustment, we try the third way that the variables are optimized in stage, which is to execute the loosely coupled bundle adjustment first, and then the tightly coupled bundle adjustment for higher accuracy.

V. PRIOR LASER MAP BUILDING AND FILTERING

For more accurate data association between laser and visual maps, we extract the salient and stable subset from multi-session laser and visual data. To begin with, we model the correspondence between the two modalities.

The localization problem against a given map can be formulated as a maximum likelihood problem:

$$
\xi^* = \arg\max\log p(\mathcal{M}^v|\mathcal{M}^l;\xi)
\tag{10}
$$

To demonstrate the data association process, the correspondence variable c_i is introduced which indicates that p_i^v

is matched with p_j^l if $c_i = j$, thus the likelihood L could be substituted as:

$$L = \sum_i \log \sum_j p(p_i^v, c_i = j | p_j^l; \xi) \tag{11}$$

Suppose $Q_i(c_i)$ as a density function of c_i, the lower bound of L can be deduced upon Jensen inequality as:

$$L \geq \sum_i \sum_j Q_i(c_i) \log \frac{p(p_i^v, c_i = j | p_j^l; \xi)}{Q_i(c_i)} \tag{12}$$

This equality only exists if $Q_i(c_i)$ is the posterior of c_i

$$Q_i(c_i) = p(c_i = j | p_i^v, p_j^l; \xi) \tag{13}$$

Similar to [26], we utilize Gaussian distribution to model the likelihood

$$p(p_i^v | c_i = j, p_j^l; \xi) = N(p_i^v; R(\xi)p_j^l + t(\xi), \sigma) \tag{14}$$

$R(\xi)$ and $t(\xi)$ represent the rotation and translation components of ξ. If there is no prior on c_i, the posterior is

$$p(c_i = j | p_i^v, p_j^l; \xi) = \frac{p(p_i^v | c_i = j, p_j^l; \xi)}{\sum_j p(p_i^v | c_i = j, p_j^l; \xi)} \tag{15}$$

The problem described in (10) can be solved with Expectation Maximization (EM) algorithm similar to [27]. If we simplify the posterior as one-peak distribution at the center of the corresponding j which is the maximum of $p(p_i^v | c_i = j, p_j^l; \xi)$, the E-step is equivalent to search the nearest neighbor (NN) of p_i^v in transformed $\mathcal{M}^l$ as correspondence, which is what we utilize in Section IV, thus further leads to the solution of EM algorithm of (10) equivalent to ICP algorithm.

To guide the EM to the correct solution, the data association of NN must be close to the correct correspondence. In another word, the posterior (15) should be similar with the shape of the finite discrete normal distribution $G(c_i)$ centering at the correct data association

$$G(c_i = j) = N(p_i^v - R(\xi^*)p_j^l - t(\xi^*); 0, \eta) \tag{16}$$

where ξ^* means the correct pose. We measure the closeness of these two distributions (15) and (16) with Kullback-Leibler divergence (KLD)

$$KLD = \sum_j p(c_i = j | p_i^v, p_j^l; \xi) \log \frac{p(c_i = j | p_i^v, p_j^l; \xi)}{G(c_i = j)} \tag{17}$$

If we ignore the candidate associations with both very small (15) and (16), to reduce (17) we need to increase $p(c_i = j^* | p_i^v, p_j^l; \xi)$ where $j^* = \arg\max G(c_i = j)$, which basically can be achieved following two ways:

- **Pull the correct correspondence**: modify $\mathcal{M}^l$ to make NN of p_i^v in the modified laser map have very large $G(c_i)$, which optimally is equivalent to the $p_{j^*}^l$.
- **Push the wrong correspondence**: modify $\mathcal{M}^l$ to make wrong candidate points in the modified laser map have very small $G(c_i)$.

Guided by these two cues, we design a map optimization module which processes the laser map following the above derived theoretical insights to yield $\mathcal{M}^{vl}$.

A. Vision-based map transformation

Firstly we "pull" the correct correspondence in each data session by selecting out $p_{j^*}^l$ that is the correct correspondence of p_i^v given the groundtruth pose ξ^*.

To achieve this, we record both the laser and visual information during data collection process. The trajectories are calculated with laser-based SLAM algorithm [28] (the first trajectory) or laser-based localization algorithm (the following trajectories), thus all sessions of collected data are represented in the same map coordinate frame $\mathcal{G}$. Transform each laser scan into its corresponding camera frame then project the laser points onto the image

$$u_j^l = \pi(R(\xi_{\mathcal{F}}^{\mathcal{C}})p_j^l + t(\xi_{\mathcal{F}}^{\mathcal{C}})) \tag{18}$$

where $\pi(\cdot)$ represents the reprojection function and $\xi_{\mathcal{F}}^{\mathcal{C}}$ means the extrinsic transformation between camera and laser coordinate frames.

We extract feature points on each image using the same type of extractor which will be used in visual inertial system. Then among the extracted feature points we search the closet point u_j^v for each u_j^l. If the distance between them satisfies

$$D(u_i^v, u_j^l) < \delta_u \tag{19}$$

where $D(\cdot, \cdot)$ returns the distance between the input points, then u_j^l is considered as a correct correspondence of u_j^v and p_j^l is inserted into the vision-transformed map of the corresponding session. When all of the collected sessions have been processed, these vision-transformed maps $\{\mathcal{M}_i^{vl}\}_{i=0}^{n-1}$ are delivered into the next statistic-filtering module for further optimization.

B. Static map maintenance

To further "push" away the wrong correspondence, we select only the stable laser points out of multi sessions of data to avoid wrong data association which could be introduced by the dynamics.

Initially, we keep all of the points in $\mathcal{M}_0^{vl}$ as the base map $\mathcal{M}^{vl}$. For each laser point $p_j^{vl, \mathcal{M}_i^{vl}}$ in map $\mathcal{M}_i^{vl}$, we search its nearest neighbor $p_k^{vl, \mathcal{M}^{vl}}$ in $\mathcal{M}^{vl}$. If the distance satisfies

$$D(p_j^{vl, M_i^{vl}}, p_k^{vl, M^{vl}}) > d^\alpha \tag{20}$$

$p_j^{vl, M_i^{vl}}$ might be collected from the new part of the environment that has never been observed in the past processed i sessions of data. Otherwise the matched point $p_k^{vl, M^{vl}}$ is considered as being observed another time.

After all of the collected sessions have been processed, the points in the merged base map $\mathcal{M}^{vl}$ should cover most part of the environment, dynamics included. We denote the number that each laser point has been observed as "observation number", which could be utilized as an index to determine whether a laser point belongs to the static part of the environment or not

$$\mathcal{M}^{vl} = \begin{cases} \mathcal{M}_s^{vl}, & obs(p_i^{vl, \mathcal{M}_s^{vl}}) \geq \beta \\ \mathcal{M}_d^{vl}, & obs(p_i^{vl, \mathcal{M}_d^{vl}}) < \beta \end{cases} \tag{21}$$

where $obs(\cdot)$ returns the observation number of the input point. $\mathcal{M}_s^{vl}$ represents the subset in $\mathcal{M}^{vl}$ in which the points belong to the static part of the environment, and $\mathcal{M}_d^{vl}$ represents the dynamic part.

Since we do not apply any model assumption when classify the map points, we further apply two post-processing steps to discover the false positive and false negative classification results which are introduced by overlapping. To filter out the false positive points in map $\mathcal{M}_s^{vl}$, we search the nearest point $p_j^{vl,\mathcal{M}_d^{vl}}$ in map $\mathcal{M}_d^{vl}$ for each point $p_i^{vl,\mathcal{M}_s^{vl}}$ then apply an erosion method to filter out those points in $\mathcal{M}_s^{vl}$ that satisfy

$$\begin{cases} D(p_i^{vl,M_s^{vl}}, p_j^{vl,M_d^{vl}}) < d^{er} \\ obs(p_i^{vl,\mathcal{M}_s^{vl}}) < \beta^{er} \end{cases} \quad (22)$$

Next, to fetch the false negative points which have been classified into map $\mathcal{M}_d^{vl}$, we search the nearest point $p_j^{vl,\mathcal{M}_s^{vl}}$ for each point $p_i^{vl,\mathcal{M}_d^{vl}}$ in map $\mathcal{M}_s^{vl}$ then apply an expansion method to select those points in $\mathcal{M}_d^{vl}$ which satisfy

$$\begin{cases} D(p_i^{vl,M_d^{vl}}, p_j^{vl,M_s^{vl}}) < d^{ep} \\ obs(p_i^{vl,\mathcal{M}_d^{vl}}) > \beta^{ep} \end{cases} \quad (23)$$

then add them back into map $\mathcal{M}_s^{vl}$. After the post-processing steps, most of the map $\mathcal{M}_s^{vl}$ should belong to the static and vision-related subset of the original laser map.

C. Ground points modification

When extracting feature points on an image, some feature points might be extracted from the textureless area or edges of shadow on the ground. These points are highly sensitive to the illumination, which means that they are hardly extracted at the same position at another time, thus are excluded from the map $\mathcal{M}_s^{vl}$. However, some of the extracted points in current session might also belong to this class. As a result, the NN based search may encourage these points to match the map, which must leads to the wrong data association, especially when the initial pose is noisy.

To improve the robustness of data association when the initial pose is noisy, we assign the ground points with the point-to-plane error term, as the ground is locally planar. The point-to-plane error term only care about the perpendicular distance, thus any data association between points on the local ground can be regarded correct. We select the laser points from the ground in each scan of data based on a height filter as the position of the LiDAR with respect to the robot is considered as known and fixed. All of the ground points are merged and further filtered by a voxel filter into the ground map. Finally, the filtered map $\mathcal{M}^{vl}$ and the sampled ground map are merged into the final cross-modal stable map $\mathcal{M}^{vl}$ for visual inertial localization.

VI. Experimental Results

To evaluate the performance of the proposed vision-related laser map optimization method and the laser map aided visual inertial localization method, we collect data in
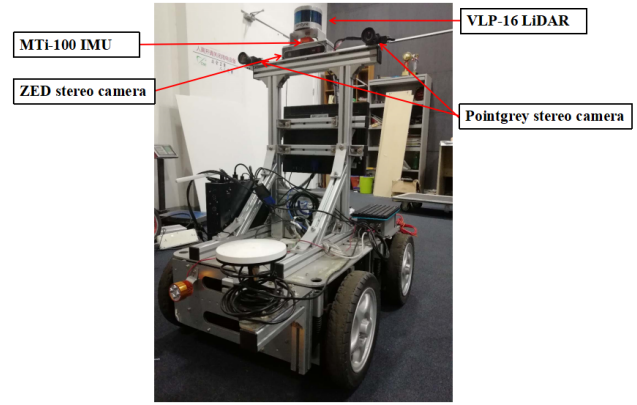

Fig. 4: The equipment used for collecting data.

TABLE I: Overview of YQ South datasets

Map Building Dataset (early spring)			
Start Time	Duration	Start Time	Duration
2017/03/03 07:52:31	17:44	2017/03/03 09:20:13	18:45
2017/03/03 10:23:11	18:14	2017/03/03 11:48:03	18:17
2017/03/03 12:59:16	19:12	2017/03/03 14:34:43	19:24
2017/03/03 16:05:54	18:39	2017/03/03 17:38:14	18:01
2017/03/07 07:43:30	17:54	2017/03/07 09:06:04	18:46
2017/03/07 10:19:45	19:04	2017/03/07 12:40:29	18:42
2017/03/07 14:35:16	19:01	2017/03/07 16:28:26	17:59
2017/03/07 17:25:06	18:34	2017/03/07 18:07:21	19:49
2017/03/09 09:06:05	17:50	2017/03/09 10:03:57	17:52
2017/03/09 11:25:40	18:17	2017/03/09 15:06:14	19:13
2017/03/09 16:31:34	19:36		
Testing Dataset (summer and winter)			
Start Time	Duration	Start Time	Duration
2017/08/23 09:40:13	16:31	2017/08/24 09:21:41	13:21
2017/08/27 15:22:11	17:03	2017/08/28 17:06:06	17:15
2018/01/29 11:09:15	14:59		

southern Yuquan Campus, Zhejiang University, China. The experimental platform to collect data for map construction is a four-wheel mobile robot as shown in Fig. 4 equipped with a VLP-16 Velodyne LiDAR, a MTi 100 IMU and a ZED stereo camera. When collect data for evaluating the localization method, the Zed is replaced by a Pointgrey stereo camera to verify the uncorrelation with equipment. The methods are deployed on a computer with an Intel i7-6700 CPU 3.40GHz and 16G RAM using C++. The ICP algorithm used in our method is implemented based on libpointmatcher [28]. We extract ORB features during localization referring to [29] and solve the nonlinear optimization problem with Ceres [30].

A. YQ datasets overview

The dataset used for map building includes 21 sessions of data collected from three days in March, 2017, which are listed detailedly in Table I. The robot is under remote control and each session of data is collected almost along the same way with the length of around 1,300 meters as shown in Fig. 1. As a campus environment, the high-dynamics include pedestrians, cyclists, moving cars and buses. While the low-dynamics include parking cars, the shape of the trees, the landmarks on the road and products of the weather like snow.

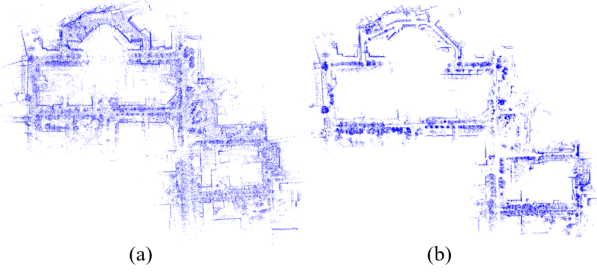The dataset for testing the visual inertial localization

Fig. 5: The full laser map (a) and the filtered laser map (b).

Sequences	tightly	1:1	3:1	5:1	loosely
23/08/2017 09:40:13	0.580	0.473	0.347	**0.315**	0.792
24/08/2017 09:21:41	0.662	0.494	0.484	**0.447**	1.015
27/08/2017 15:22:11	0.956	**0.417**	0.468	0.480	0.944
28/08/2017 17:06:06	0.582	0.434	**0.392**	0.440	0.721
29/01/2018 11:09:15	0.683	0.429	0.435	**0.391**	0.605
average	0.693	0.449	0.425	**0.415**	0.815

method contains four sessions of data collected in summer and one session of data collected in winter as shown in Table I, in which a session of summer data (2017/08/24) is collected along the opposite way of the others to test the bi-directional performance. Besides, the other three sessions of data are collected at the different time of a day. And the last session of data is collected after snow where the snow extremely changes the visual features on the road.

We utilize the laser-based localization method [28] to evaluate the trajectories for each session of data as the groundtruth. The map used for laser-based localization is constructed using the firstly collected session of data. Thus all of the groundtruth is represented in the same coordinate.

B. Map construction

We utilize 21 sessions of data collected during three days in March, 2017. All of the sessions are merged using the method in Section V. The final extracted map is shown in Fig. 5 (b). The map shown in Fig. 5 (a) is constructed with laser-based SLAM method [31] using the first session of data. It's easy to figure out that the spatial distributions of these two maps are different and the spatial distribution of the filtered map is more like the one of traditional visual constructed map.

C. Evaluations of localization

To validate the efficiency of the proposed localization method, we evaluate it in the five sessions of data collected in summer and winter. We apply the hybrid bundle adjustment with different frequency of the tightly coupled bundle adjustment. The frequency is demonstrated as the proportion of executed times. For example, "3:1" means after three times of loosely coupled bundle adjustment one time of tightly coupled bundle adjustment will be executed. Also the performance of totally applying the tightly coupled bundle adjustment and loosely coupled bundle adjustment are also evaluated. The results are shown in Table II, inside the bold numbers represent the best results for each session of data, and the corresponding trajectories as well as the lateral and heading errors are shown in Fig. 6.

As the results show, the hybrid method outperforms the pure tightly coupled and loosely coupled methods in every session of data. On average, the proportion "5:1" gives the best performance.
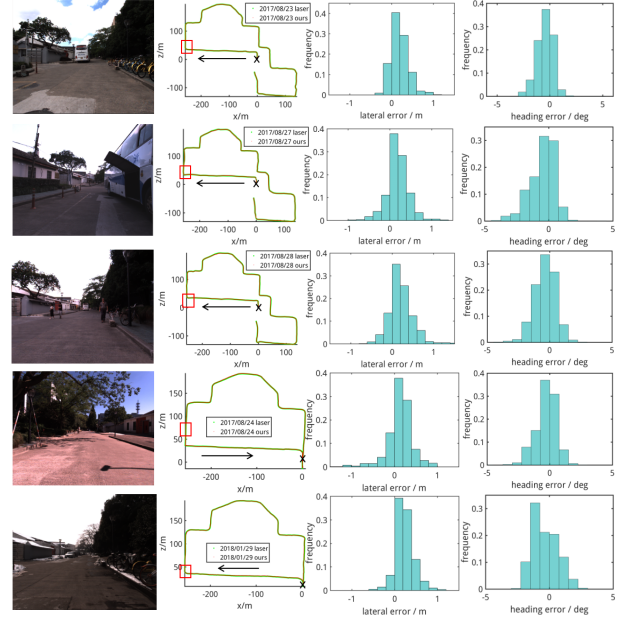

Fig. 6: Localization trajectories (second column) and histograms of lateral (third column) and heading (fourth column) errors in YQ South datasets. The black crosses indicate the beginning of the trajectories and the arrows indicate the directions of the trajectories. The pictures in the first column show the views observed in different sessions at the positions that labeled in the second column with red rectangles.

D. Evaluations of map optimization

For comparison, we also evaluate the performance of localizing within the full map that is not filtered by the proposed map optimization method. The results are listed in Table III. The hybrid method for full map aided localization utilizes the same frequency at which the filtered map aided method gives the best results.

From the table, we can see that in some sequences, the results of full map aided method are almost the same with the filtered map aided method. But in other sequences their results differ a lot, especially in the session collected in 23, Aug, 2017 where the loosely coupled bundle adjustment based method fails to run over the whole trajectory within the full map. This validates the efficiency of the map optimization method for long-term localization tasks.

TABLE III: comparison of the localization performance in different maps

Sequences	filtered map		full map	
	hybrid	loosely	hybrid	loosely
23/08/2017 09:40:13	**0.315**	0.792	0.847	/
24/08/2017 09:21:41	**0.447**	1.015	0.459	0.451
27/08/2017 15:22:11	**0.417**	0.944	0.509	0.564
28/08/2017 17:06:06	**0.392**	0.721	1.879	0.539
29/01/2018 11:09:15	**0.391**	0.605	0.536	0.408
average	**0.415**	0.815	0.843	0.488

VII. Conclusions

In this paper, the cross-modal localization problem is addressed based on a hybrid bundle adjustment framework by aligning the sparse local visual submap to a LiDAR built map. Both the uncertainties of local visual submap and localization estimation are optimized simultaneously. To improve the accuracy of data association between the sparse visual map and dense laser map, a saliency map optimization method is proposed which also filters out the dynamics in the laser data. The efficiency of this proposed method is validated with experiments on multiple sessions of real-world data, which include the cross-seasonal and bi-directional circumstances.

In the future, we would like to learn the map extraction method proposed in this paper to filter the laser maps of the new environment. In addition, the visual SLAM aided by the prior laser map is also our potential research direction.

Acknowledgement

References

[1] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[2] M. Paton, F. Pomerleau, K. MacTavish, C. J. Ostafew, and T. D. Barfoot, "Expanding the limits of vision-based localization for long-term route-following autonomy," *Journal of Field Robotics*, vol. 34, no. 1, pp. 98–122, 2017.

[3] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[4] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2016, pp. 1926–1931.

[5] A. Gawel, T. Cieslewski, R. Dub, M. Bosse, R. Siegwart, and J. Nieto, "Structure-based vision-laser matching," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2016, pp. 182–188.

[6] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 176–183.

[7] P. Neubert, S. Schubert, and P. Protzel, "Sampling-based methods for visual navigation in 3d maps by synthesizing depth images," in *Intelligent Robots and Systems (IROS 2017), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 2492–2498.

[8] G. Pascoe, W. Maddern, and P. Newman, "Robust direct visual localisation using normalised information distance," in *British Machine Vision Conference*, 2015, pp. 70.1–70.13.

[9] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp." in *Robotics: science and systems*, vol. 2, no. 4, 2009, p. 435.

[10] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Robotics and automation, 2007 IEEE international conference on*. IEEE, 2007, pp. 3565–3572.

[11] K. J. Wu, A. M. Ahmed, G. A. Georgiou, and S. I. Roumeliotis, "A square root inverse filter for efficient vision-aided inertial navigation on mobile devices," in *2015 Robotics: Science and Systems Conference, RSS 2015*. MIT Press Journals, 2015.

[12] 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*, vol. 34, no. 3, pp. 314–334, 2015.

[13] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2017.

[14] T. Schneider, M. T. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An open framework for research in visual-inertial mapping and localization," *IEEE Robotics and Automation Letters*, 2018.

[15] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, 2015.

[16] A. Wendel, A. Irschara, and H. Bischof, "Natural landmark-based monocular localization for mavs," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5792–5799.

[17] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular slam on a mobile phone," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 4, pp. 531–539, 2014.

[18] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[19] M. Dymczyk, S. Lynen, T. Cieslewski, M. Bosse, R. Siegwart, and P. Furgale, "The gist of maps-summarizing experience for lifelong localization," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2767–2773.

[20] 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*, vol. 32, no. 6, pp. 1309–1332, 2016.

[21] P. Ozog, M. Johnson-Roberson, and R. M. Eustice, "Mapping underwater ship hulls using a model-assisted bundle adjustment framework," *Robotics and Autonomous Systems*, vol. 87, 2016.

[22] J. McDonald, M. Kaess, C. Cadena, J. Neira, and J. J. Leonard, "6-dof multi-session visual slam using anchor nodes," 2011.

[23] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[24] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.

[25] S. Wright and J. Nocedal, "Numerical optimization," *Springer Science*, vol. 35, no. 67-68, p. 7, 1999.

[26] Y. Wang, R. Xiong, and Q. Li, "Em-based point to plane icp for 3d simultaneous localization and mapping," *Int J Rob Autom*, vol. 28, pp. 234–244, 2013.

[27] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1722–1729.

[28] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, "Comparing icp variants on real-world data sets," *Autonomous Robots*, vol. 34, no. 3, pp. 133–148, April 2013.

[29] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[30] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org.

[31] F. Pomerleau, P. Krüsi, F. Colas, P. Furgale, and R. Siegwart, "Long-term 3d map maintenance in dynamic environments," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3712–3719.