Globally optimal consensus maximization for robust visual inertial localization in point and line map

Yanmei Jiao[1], Yue Wang[1], Bo Fu[1], Qimeng Tan[2], Lei Chen[2], Shoudong Huang[3] and Rong Xiong[1]

***Abstract*— Map based visual inertial localization is a crucial step to reduce the drift in state estimation of mobile robots. The underlying problem for localization is to estimate the pose from a set of 3D-2D feature correspondences, of which the main challenge is the presence of outliers, especially in changing environment. In this paper, we propose a robust solution based on efficient global optimization of the consensus maximization problem, which is insensitive to high percentage of outliers. We first introduce *translation invariant measurements* (TIMs) for both points and lines to decouple the consensus maximization problem into rotation and translation subproblems, allowing for a two-stage solver with reduced solution dimensions. Then we show that (i) the rotation can be calculated by minimizing TIMs using only *1-dimensional branch-and-bound* (BnB), (ii) the translation can be found by running 1-dimensional search for three times with *prioritized progressive voting*. Compared with the popular randomized solver, our solver achieves deterministic global convergence without depending on an initial value. While compared with existing BnB based methods, ours is exponentially faster. Finally, by evaluating the performance on both simulation and real-world datasets, our approach gives accurate pose even when there are 90% outliers (only 2 inliers).**

I. Introduction

Visual inertial navigation is currently a popular option for state estimation in mobile robots, autonomous vehicles and augmented reality applications. Many efforts have been paid to build accurate, consistent and efficient visual inertial odometry [1] [2]. However, its inherent drift is unacceptable in long-term operation, calling for absolute pose estimation for correction. Map based visual inertial localization is therefore an important component in a complete navigation system, of which the underlying problem is to estimate the absolute pose from a set of feature correspondences between 2D image key points and global 3D map points. In this problem, one main challenge is the robustness of the solver against the outliers (incorrect feature correspondences). When high percentage of correspondences is outlier, the performance of the general pose estimator may seriously degenerate.

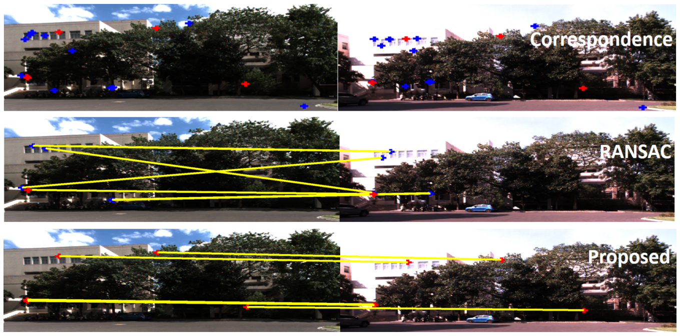


Fig. 1: The projected map points on the map image (left column) and the detected image key points on the query image (right column), with inlier correspondences in red and outliers in blue. The initial correspondences found by feature descriptor matching (top), and the consensus set correspondences searched by RANSAC (middle) and proposed consensus maximization algorithm (bottom).

Pose estimation with outliers is in general stated as consensus maximization problem. One popular solution is random sample consensus (RANSAC), which has lots of variants [3] [4] and has been employed in many visual localization methods [5] [6]. The advantage of RANSAC is the simplicity for implementation, and the usefulness in many scenarios with moderate percentage of outliers. But there are also disadvantages that (i) it cannot tolerate extreme percentage of outliers, say 90%, (ii) it is a probabilistic method, thus not guaranteeing the deterministic global optimality.

In contrast to RANSAC, another solution to consensus maximization is global optimization based methods, which can give globally optimal solution without referring to an initial value [7] [8]. However, one obstacle preventing its application is the considerable computation time. Most global optimization methods aim at general geometry estimation problems. They employ branch-and-bound (BnB) as the basic framework to reduce the search space [9], or mixed integer programming for further acceleration [10] [11]. But the computational cost is still unsatisfactory as the multi-dimensional search space $SE(3)$ is coupled.

In this paper, we propose a deterministic visual inertial localization solution to achieve global convergence with much higher efficiency by dividing $SE(3)$ search space into multiple 1-D search spaces. Specifically, inspired by the minimal solution in RANSAC, we build intermediate cost function for both point and line features, *translation invariant measurements* (TIMs), to decouple consensus maximization into two cascaded subproblems only related to rotation $SO(3)$ and translation $\mathbb{R}^3$ respectively. Based on TIMs, the globally optimal rotation is then searched by *1-dimensional BnB* in $[-\pi, \pi]$ with the aid of inertial measurements. For the translation part, $\mathbb{R}^3$ search is replaced with three times

[1] Yanmei Jiao, Yue Wang, Bo Fu, Rong Xiong are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, P.R. China. [2] Qimeng Tan and Lei Chen are with the Beijing Key Laboratory of Intelligent Space Robotic System Technology and Applications, Beijing Institute of Spacecraft System Engineering, Beijing, P.R. China. [3] Shoudong Huang is with the Center for Autonomous Systems (CAS), the University of Technology, Sydney, Australia. Yue Wang is the corresponding author `wangyue@iipc.zju.edu.cn`.

1-dimensional $\mathbb{R}$ search using *prioritized progressive voting*. To the best of our knowledge, this is the first solver for visual inertial localization with deterministic global optimality. In summary, the contributions include

- TIMs based formulation of visual inertial localization decouples the problem and enables 1D BnB based global optimization of the rotation.
- Prioritized progressive voting method replaces $\mathbb{R}^3$ space search with three times $\mathbb{R}$ search for global optimization of the translation.
- Experiments on simulation and real-world cross-session datasets that validate the effectiveness and efficiency of the proposed method against comparative methods.

The remainder of the paper is organized as follows: Section II reviews the related literatures. Section III presents the decoupling of the consensus maximization problem. Section IV introduces the solutions of the subproblems. Section V demonstrates the experimental settings and results, followed by Section VI concluding the paper.

II. Related Works

A. Visual localization

Visual localization and navigation for mobile robots has been studied extensively in the robotics and computer vision communities in the recent decade. A general visual navigation system has two components: visual odometry, which estimates the relative pose and has drift in long term [12] [13], and visual localization, which eliminates the drift by registering the image on a global map [14] [15]. More recently, inertial sensors are employed in the system to improve the accuracy and robustness [16] [17] [18]. Specifically, the inertial sensor has globally observable pitch and roll measurements, reducing the degrees of freedom (DoF) in visual inertial localization problem to 4. In [1] [19], the reduction is utilized when formulating the pose estimation given a set of inlier feature correspondences. However, few works have been done on outliers elimination when inertial measurements are provided.

B. Random sample consensus

For robust localization given the feature correspondences containing outliers, RANSAC is the most popular solution employed in many visual navigation system. To deal with the visual localization without inertial measurements, i.e. 6DoF, there have been many variants. In [20] [21] [22], point feature correspondences based RANSAC are studied. In [23] [24] [25], RANSAC is extended to line features. When inertial measurements are provided, the DoF of the problem is reduced, which is utilized by RANSAC to improve the robustness in [26] [27], and extended to both point and line correspondences in [28]. As RANSAC is developed on randomized sampling theory, it is simple to implement and has good performance on scenarios with moderate outliers. But its disadvantage is also obvious, including low tolerance against extreme outliers, local convergence and no guarantee of the optimality [29].

C. Global optimization method

Global optimization methods are proposed to achieve the global optimality and deterministic convergence, addressing the shortcomings of RANSAC. In this branch of literatures, Branch-and-Bound (BnB) is mostly used, which gradually prunes the solution space by coarse-to-fine division. In [30], BnB is used to solve the 2D-2D registration problems. In [9], a general framework for point, line and plane features is proposed to solve 3D-3D registration via BnB. Integrated with mixed integer programming, the BnB optimization can converge faster [10] [11]. In [29], the linear matrix inequality constraints are introduced to mixed integer programming, resulting in a faster BnB for all 2D-2D, 2D-3D and 3D-3D geometric vision problems. In the works mentioned above, the rotation is modeled as a rotation matrix with matrix level constraints. Thus it is unclear about the incorporation of inertial measurements. In addition, there are also works propose globally optimal algorithms specializing on one class of problem. In [31] [32], pairs of features are used to decouple the 3D-3D registration. In [33], TEASER is proposed for decoupled scaled 3D-3D registration. These works show that it is possible to have superior performance with specialized algorithms rather than only the general BnB framework, even accelerated.

In this paper, we follow the idea of specialized solver to bridge the gap of globally optimal deterministic solution for visual inertial localization, which is a robust 3D-2D pose estimation problem with inertial measurements. To the best of our knowledge, this is the first work to study this problem in the context of global optimality. We expect this solution to be accurate and efficient.

III. Decoupling Translation and Rotation

The underlying problem of visual inertial localization is the pose estimation from 3D-2D correspondences with outliers. Formally, given a set $\mathfrak{P}$ consisting of correspondences between 3D global points $p_i \in \mathbb{R}^3$ and 2D visual points $u_i \in \mathbb{R}^2$, they satisfy

$$u_i = \pi(Rp_i + t, K) + o_i + e_i \tag{1}$$

where $R \in SO(3)$ and $t \in \mathbb{R}^3$ is the camera pose to be estimated, π is the camera projection function with known intrinsic parameters K, $|e_i| < n_i$ is assumed to be bounded random measurement noise, o_i is zero for inlier while an arbitrary number for outlier. To deal with outliers, the robust pose estimation generally begins with consensus maximization problem as

$$\max_{R,t,\{z_i\}} \sum z_i \tag{2}$$

$$s.t. \quad z_i |u_i - \pi(Rp_i + t, K)| \leq n_i, \quad i \in \mathfrak{P} \tag{3}$$

where z_i is binary, indicating whether o_i is zero. To solve the problem in global, general BnB algorithms search in $SE(3)$, which is a coupled space of $SO(3)$ and $\mathbb{R}^3$. But this probably leads to exponential computational complexity in bad cases. For local techniques like RANSAC, inliers may be estimated conservatively, i.e. inliers regarded as outliers, especially when the noise is unavoidable.

A. Translation invariant measurements

1) Point-TIM: Inspired by the minimal solution in RANSAC, we develop an intermediate measurement which is invariant to the translation of the pose. Mathematically, given an image key point u_i, we have an un-normalized direction vector from the camera center as

$$\tilde{u}_i \triangleq \begin{pmatrix} \tilde{u}_{i,x} \\ \tilde{u}_{i,y} \\ 1 \end{pmatrix} = K^{-1} \begin{pmatrix} u_i \\ 1 \end{pmatrix} \tag{4}$$

Then the corresponding world point p_i is transformed to the camera coordinates and satisfies

$$\frac{R_1 p_i + t_x}{\tilde{u}_{i,x}} = \frac{R_2 p_i + t_y}{\tilde{u}_{i,y}} = R_3 p_i + t_z \tag{5}$$

where $R \triangleq (R_1^T, R_2^T, R_3^T)^T$ and $t \triangleq (t_x, t_y, t_z)^T$. Based on (30), we have two constraints from a correspondence. Naturally, given another correspondence u_j and p_j, we can have two more constraints as

$$\frac{R_1 p_j + t_x}{\tilde{u}_{j,x}} = \frac{R_2 p_j + t_y}{\tilde{u}_{j,y}} = R_3 p_j + t_z \tag{6}$$

According to (30) and (33), we have linear constraints of the translation t. With proper variable substitutions among the constraints, and the globally observable pitch and roll angles from inertial measurements, we can eliminate t, reduce $SO(3)$ to $[-\pi, \pi]$, and derive TIM as

$$d_p(\alpha) = d_{p,1} \sin\alpha + d_{p,2} \cos\alpha + d_{p,3} \tag{7}$$

where α is the unknown yaw angle, $d_{p,1}$, $d_{p,2}$, $d_{p,3}$ and the derivation details are presented in the Appendix. Now we substitute the constraints which are related to both R and t in (2) with the TIM, leading to

$$\max_{R(\alpha),\{z_{ij}\}} \sum z_{ij} \tag{8}$$

$$s.t. \quad z_{ij} |d_{p,ij}(\alpha)| \leq n_{ij}, \quad i,j \in \mathfrak{P} \tag{9}$$

where $n_{ij} = \min(n_i, n_j)$, $z_{ij} = 1$ indicates the i-th and j-th correspondence derived the constraint are inliers.

2) Line-TIM: Similar to a pair of point correspondences, given a set of line correspondences $\mathfrak{L}$, it is also possible to develop TIM. Given the end points of the image line segment u_{k1} and u_{k2}, we have two un-normalized directions as (29), denoted as $\tilde{u}_{k1}$ and $\tilde{u}_{k2}$.

Then following the fact that the point p_i on the world line lies on the plane spanned by the rays from camera center along direction $\tilde{u}_{k1}$ and $\tilde{u}_{k2}$, we have

$$(\tilde{u}_{k1} \times \tilde{u}_{k2})^T (R p_k + t) = 0 \tag{10}$$

which is a constraint for both rotation and translation. Since arbitrary number of points can be sampled from a line, we sample another point on the same world line to formulate the constraint as (10). Then only one line correspondence can lead to line-TIM after proper substitution as

$$d_l(\alpha) = d_{l,1} \sin\alpha + d_{l,2} \cos\alpha + d_{l,3} \tag{11}$$

where the line-TIM has the same form as point-TIM in (36), but the coefficients are different. The derivation details are also presented in the Appendix.

TIMs based rotation only problem. Note that either (36) or (42) is only related to the yaw angle. By combining them together, we have a general consensus maximization problem with TIM constraints only relating to rotation compatible to the map having both point and line features as

$$\max_{R(\alpha),\{z_*\}} \sum z_* \tag{12}$$

$$s.t. \quad z_{ij} |d_{p,ij}(\alpha)| \leq n_{ij}, \quad i,j \in \mathfrak{P} \tag{13}$$

$$z_k |d_{l,k}(\alpha)| \leq n_k, \quad k \in \mathfrak{L} \tag{14}$$

B. Two-stage consensus maximization solver

With TIMs for both point and line correspondences, we decouple the original consensus maximization problem into rotation only problem, and translation only problem when the rotation is fixed. Accordingly, the proposed solver has two stages in cascade:

- We estimate the rotation $\hat{R}$ by $R(\hat{\alpha})$ based on the TIMs in (12). This estimator solves a 1D optimization problem and is described in Section IV-A.
- We estimate the translation $\hat{t}$ based on the original consensus maximization in (2) where the rotation is assigned with $\hat{R}$. This estimator solves a $\mathbb{R}^3$ optimization problem and is described in Section IV-B.

IV. Estimators of Rotation and Translation

A. BnB based optimization for rotation

We employ BnB strategy to solve problem (12). The cost function in (12) relates to α and z_*. But it is obvious that when α is determined, $\{z_*\}$ is simply derived by evaluating the constraints. So we denote the cost function as $E(\alpha)$ that is explained as the number of inliers given a yaw angle α.

Upper bound of cost function. We then derive the upper bound of $E(\alpha)$ on the subset $\mathbb{A}$, denoted as $\overline{E}(\mathbb{A})$, where $\alpha \in \mathbb{A} \subseteq [-\pi, \pi]$. Recall (36) and (42), as the forms of point-TIM and line-TIM are the same, we denote them as $d(\alpha)$. The lower bound of $|d(\alpha)|$ on $\mathbb{A}$, denoted as $\underline{d}(\mathbb{A})$, is derived as

$$\underline{d}(\mathbb{A}) = \min |a_1 \sin(\alpha + a_2) + d_3| \tag{15}$$

where the derivation of the coefficients are introduced in Appendix. Note that $\underline{d}(\mathbb{A})$ can be solved analytically without any iterations. Now we formulate a consensus maximization problem as

$$\max_{R(\alpha),\{z_*\},\alpha\in\mathbb{A}} \sum z_* \tag{16}$$

$$s.t. \quad z_{ij} \underline{d}_{p,ij}(\mathbb{A}) \leq n_{ij}, \quad i,j \in \mathfrak{P} \tag{17}$$

$$z_k \underline{d}_{l,k}(\mathbb{A}) \leq n_k, \quad k \in \mathfrak{L} \tag{18}$$

where the problem is defined on $\mathbb{A}$, and the TIMs constraints are replaced with tight lower bounds, relaxing the constraints and yielding an optimistic estimation of $\hat{z}_*$. We then have

$$E(\alpha) \leq \overline{E}(\mathbb{A}) = \sum \hat{z}_*, \quad \alpha \in \mathbb{A} \tag{19}$$

as a tight upper bound. The equality exists when all constraints give the same α with $d_{p,ij}(\alpha) = \underline{d}_{p,ij}(\mathbb{A})$ and $d_{l,k}(\alpha) = \underline{d}_{l,k}(\mathbb{A})$, which is only possible when noise is free.

Accelerate BnB optimization. With (12-19), we have the BnB search for globally optimal rotation, of which the pseudo code is listed in Algorithm 1. Note that the main idea of BnB is to prune the solution space $\mathbb{A}$ when its upper bound $\overline{E}(\mathbb{A})$ is smaller than the current best estimates E^*. Therefore, if we have a fast solution to initialize a good E^*, most solution spaces can be pruned at early stage, significantly improving the search efficiency. To implement this idea, we use RANSAC [28] to generate a rough initial E^*. In addition, we introduce a heuristics to balance the global optimality and the efficiency. The best M estimated α during RANSAC is utilized to initialize M subsets among $[-\pi, \pi]$. Each subset centers at each estimated α with a width w. When w is large, global optimality is emphasized and vice versa. Another implementation trick is to store the respective inliers when evaluating (16) on each subset $\mathbb{A}$. When $\mathbb{A}$ is further divided into smaller subsets, only the stored inliers within $\mathbb{A}$ are evaluated, instead of all constraints, saving lots of computational cost. These techniques are all shown to accelerate the search in the experimental ablation study without drop of accuracy.

Algorithm 1: Globally Optimal Rotation Search

```
Input: 3D-2D feature correspondences $\mathfrak{P}$, $\mathfrak{L}$
Output: Optimal $\alpha^*$
Initialize partition of $[-\pi, \pi]$ into subsets $\{\mathbb{A}_i\}$.
Initialize best estimation $E^*$, $\alpha^*$.
Insert $\{\mathbb{A}_i\}$ into queue $q$.
while $q$ is not empty do
    Pop the first subset of $q$ as $\mathbb{A}$.
    Compute $\overline{E}(\mathbb{A})$ as (16).
    if $\overline{E}(\mathbb{A}) > E^*$ then
        Assign center of $\mathbb{A}$ as $\alpha_c$.
        Compute $E(\alpha_c)$ as (12).
        if $E(\alpha_c) > E^*$ then
            Update $E^* \leftarrow E(\alpha_c)$, $\alpha^* \leftarrow \alpha_c$.
        Subdivide $\mathbb{A}$ into subsets and insert into $q$.
```

B. Prioritized progressive voting for translation

When $R(\hat{\alpha})$ is estimated, the co-linear and co-planar constraints (30) and (10) are all linear constraints for t. Thus we can transform the consensus maximization problem with point and line constraints as

$$\max_{t,\{z_i\}} \sum z_i \tag{20}$$

$$s.t. \quad z_i |A_i t + b_i| \leq n_i, \quad i \in \mathfrak{P} \cup \mathfrak{L} \tag{21}$$

where $A_i \in \mathbb{R}^{1\times 3}$ and $b_i \in \mathbb{R}$ are the coefficients for linear constraints derived from (30) or (10) with estimated $R(\hat{\alpha})$. However, this problem still has coupled constraints for t so that $\mathbb{R}^3$ search is indispensable.

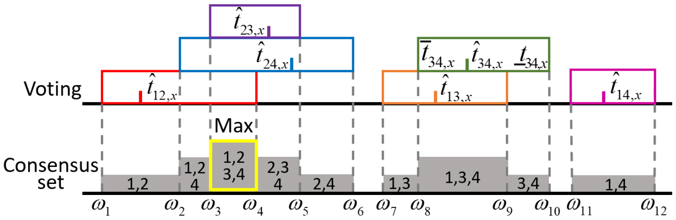


Fig. 2: The voting illustration of $\hat{t}_x$. Each $\hat{t}_{ij,x}$ derived by i-th and j-th correspondence votes for the interval if $[\omega_i, \omega_{i+1}] \subseteq [\underline{t}_{ij,x}, \overline{t}_{ij,x}]$, which means the corresponding consensus set contains i and j.

Decoupled linear constraints. Note that for a point correspondence constraint (30), we have two linear equations, while for a line correspondence constraint (10), we have one. Therefore, given a pair of correspondences including at least one point correspondence, say the i-th point correspondence and the j-th point or line correspondence, it is sufficient to solve $\hat{t}_{ij}$ for this small linear system (see Appendix for details), then we have

$$\max_{t,\{z_{ij}\}} \sum z_{ij} \tag{22}$$

$$s.t. \quad z_{ij}|\hat{t}_{ij} - t| \leq n_{ij}, \quad i \in \mathfrak{P}, j \in \mathfrak{P} \cup \mathfrak{L} \tag{23}$$

Now we find that the constraints are decoupled for each dimension of t. Set the x-dimension as example, we have

$$\max_{t_x,\{z_{ij}\}} \sum z_{ij} \tag{24}$$

$$s.t. \quad z_{ij}|\hat{t}_{ij,x} - t_x| \leq n_{ij,x}, \quad i \in \mathfrak{P}, j \in \mathfrak{P} \cup \mathfrak{L} \tag{25}$$

arriving at the resultant three dimension-wise linear constrained consensus maximization problems.

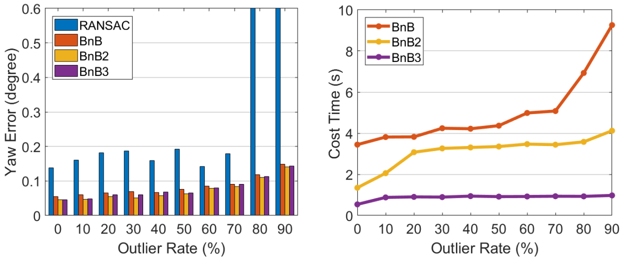
Dimension-wise voting algorithm. We use a voting algorithm to solve the problem. We first specify the noise bound $n_{ij,x}$ in (24). Given the noise bound n_i in (20), we have the noise bound for t following the techniques in [34] [35] as

$$\underline{t}_{ij} \leq \hat{t}_{ij} \leq \overline{t}_{ij} \tag{26}$$

The details can be found in Appendix.

Still taking x-dimension as example, each estimated $\hat{t}_{ij,x}$ defines an interval $[\underline{t}_{ij,x}, \overline{t}_{ij,x}]$. If the real t_x lies in this interval, then the real inlier set contains the two correspondences deriving $\hat{t}_{ij,x}$. According to [33], the insight is that the inlier set only changes its membership when real t_x enters a new interval. Besides, given K estimations, the *maximum number of possible consensus sets, i.e. the cardinality of the solution space, is* $2K-1$, where K is in *quadratic* w.r.t the number of correspondences. This complexity enables a voting algorithm for all $2K-1$ sets. By counting the unique correspondences of the votes in each set, we get the corresponding consensus set. Then the maximal consensus set can lead to an estimation of $\hat{t}_x$. An illustrative case is shown in Fig. 2 and the pseudo code is listed in Algorithm 2 with x-dimension as example. For simplicity, we replace $\hat{t}_{ij,x}$ with $\hat{t}_{k,x}$ in the pseudo code. Following the similar idea in [33], by repeating the voting algorithm for three times, $\hat{t}$ is estimated as $[\hat{t}_x, \hat{t}_y, \hat{t}_z]^T$.

Prioritized progressive voting algorithm. When the



(a) (b)

Fig. 3: The rotation accuracy and computation time over the increasing outlier rate. *BnB2* denotes the BnB with RANSAC initialization. *BnB3* denotes the BnB with both RANSAC initialization and the implementation trick.

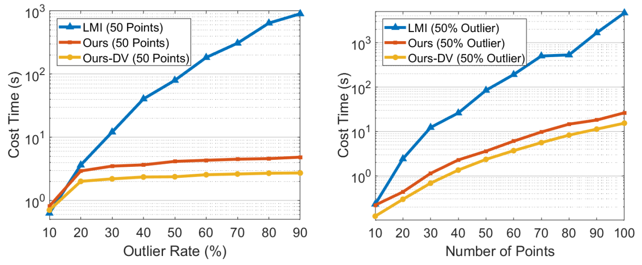
```
Algorithm 2: Voting
  Input: $\{\hat{t}_{k,x}\}$, $\{\underline{t}_{k,x}\}$, $\{\bar{t}_{k,x}\}, k = 1..K$
  Output: Consensus sets $S$
  Initialize key-value map $S$.
  $\omega = sort([\underline{t}_{1,x}, \bar{t}_{1,x}, \underline{t}_{2,x}, \bar{t}_{2,x}, .., \underline{t}_{K,x}, \bar{t}_{K,x}])$.
  for $i = 1..2K-1$ do
      $S([\omega_i, \omega_{i+1}]) = \emptyset$.
      for $k = 1..K$ do
          if $[\omega_i, \omega_{i+1}] \subseteq [\underline{t}_{k,x}, \bar{t}_{k,x}]$ then
              $S([\omega_i, \omega_{i+1}]) = S([\omega_i, \omega_{i+1}]) \cup k$.
```

number of inliers is high, independent voting along three dimensions is possible. But when the number of inliers is low and outlier rate is high, independent dimension-wise voting may lead to failure. The reason is that, though it is almost impossible that there are more outliers than inliers having the similar t, *it is possible that there are more outliers than inliers having the similar* t_x. In such scenario, search along x-dimension leads to incorrect $\hat{t}_x$, which cannot be corrected in the successive voting along y or z-dimension.

To deal with such scenario while keeping a low computational complexity, we propose a prioritized progressive voting for translation in Algorithm 3. The main idea is that we progressively vote on the three dimensions, but there is a priority, i.e. number of votes, for early termination. The experimental results show that the computational complexity of prioritized progressive voting is almost similar to the dimension-wise voting. Otherwise, it is also possible to use 3D BnB translation search for better accuracy, but it is slower because of the coupled multi-dimensional solution space. Finally, we apply nonlinear refinement to achieve the best accuracy when the maximum consensus set is found.

V. Experimental Results

In the experiments, we evaluate the proposed consensus maximization solver on (i) the feasibility and effectiveness of the subproblem solvers, (ii) the accuracy and robustness compared with existing methods, and (iii) the performance in real world visual inertial localization applications. We implement the proposed solver in MATLAB on a desktop with CPU Intel i7-7700 3.60GHz and 8G RAM.



(a) (b)

Fig. 4: Computation time comparison over increasing (a) outlier rate (b) number of points. *Ours* denotes the proposed method with prioritized progressive voting, while *Ours-DV* denotes the dimension-wise voting.

```
Algorithm 3: Prioritized Progressive Voting
  Input: $\{\hat{t}_k\}$, $\{\underline{t}_k\}$, $\{\bar{t}_k\}$, $k = 1..K$
  Output: Maximum consensus set $\hat{t}$
  Initialize best estimation $E^* = 0$.
  $S_x = Voting(\{\hat{t}_{k,x}\}, \{\underline{t}_{k,x}\}, \{\bar{t}_{k,x}\})$.
  Sort $S_x$ in decreasing cardinality.
  for each key $[i]$ in $S_x$ do
      if $|S_x([i])| < E^*$ then
          break;
      $S_y = Voting(\{\hat{t}_{k,y}\}, \{\underline{t}_{k,y}\}, \{\bar{t}_{k,y}\}, k \in S_x([i]))$.
      for each key $[j]$ in $S_y$ do
          if $|S_y([j])| < E^*$ then
              break;
          $S_z = Voting(\{\hat{t}_{k,z}\}, \{\underline{t}_{k,z}\}, \{\bar{t}_{k,z}\}, k \in S_y([j]))$.
          if $\max_{S_z([m])} |S_z([m])| > E^*$ then
              Update $E^* \leftarrow \max_{S_z([m])} |S_z([m])|$.
              Update $S^* \leftarrow \arg\max_{S_z([m])} |S_z([m])|$.
```

A. Ablation study

We build the synthetic world consisting of 3D points and lines in the cube $[-1, 1]^3$. The 2D image projections are generated with randomly sampled camera poses in $[-2, 2]^3 \times [-\pi, \pi]^3$, as well as their inlier correspondences. All the projected 2D image points are added with bounded random noise e_i with the bound $n_i = 2$. Each outlier correspondence is generated from other randomly sampled camera pose different to ground truth pose. The total number of correspondences is fixed as 50. Specifically, there are 50 point correspondences when evaluating point only methods, while 25 point and 25 line correspondences for the point and line methods. We vary the outlier percentage from 10% to 90% with a step of 10%. Statistic performance indicators are evaluated with an average of 100 Monte Carlo runs. Denoting the ground truth pose as $[R_{gt}|t_{gt}]$, we compute the translation error as $\triangle T = |\hat{t} - t_{gt}|$ in meter and the rotation error as

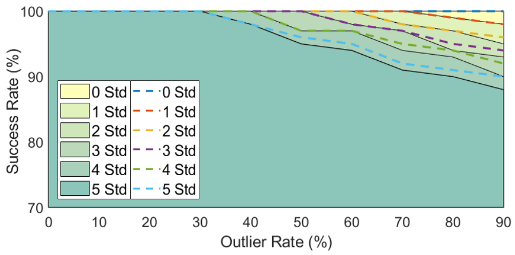


Fig. 5: The sensitivity experiment result using proposed algorithm with dimension-wise voting (solid) and prioritized progressive voting (dash).

the angle of $\triangle R = \hat{R}R_{gt}^T$ in degree.

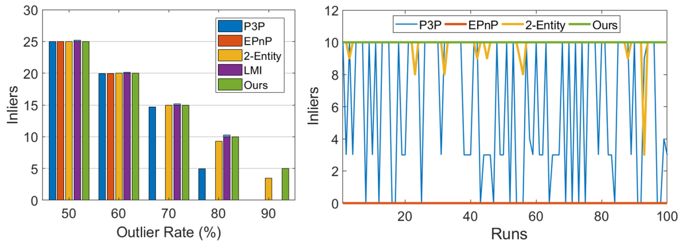
BnB heuristics. We first evaluate the heuristics introduced in Section IV-A from the aspect of accuracy and efficiency. As shown in Fig. 3, with the heuristics, the efficiency is improved while the accuracy stays similar. Since the final pose is refined by nonlinear optimization, slight rotation error after BnB can be ignored. As a baseline, we also show the error of estimated rotation giving the most inliers in RANSAC, of which the performance is much worse, indicating inconsistency between the identified inliers and the real inliers. In following experiments, heuristics are applied with BnB as default setting.

Translation voting. We then compare the voting strategies introduced in Section IV-B. Now we can evaluate the final accuracy after nonlinear refinement. In addition to efficiency and accuracy, we also evaluate the consistency between the estimated consensus set and the real inlier set (CCI) using precision and recall. As shown in Fig. 4, the computation of the prioritized progressive voting is slightly higher than the dimension-wise voting. More importantly, the increased time keeps almost consistent w.r.t outlier rate and correspondences number, which might be explained as no complexity growth for prioritized progressive voting. The CCI and accuracy are shown in the right columns in Tab. I. We see that all variants achieve perfect CCI, naturally leading to high accuracy.

Sensitivity to noisy inertial measurements. As inertial measurements are noisy, it is necessary to evaluate the sensitivity of the proposed method. We add Gaussian noise with zero mean and increasing standard deviation up to 5 degree on both pitch and roll angle. The threshold to judge a successful localization is 0.1m for translation error and 0.5 degree for rotation error as in [36]. The result is shown in Fig. 5, indicating that the proposed algorithm can achieve over 90% success rate when the noise increases to 5 degree. This level of noise is far more than the pitch and roll estimations in practice [37]. In addition, we can find that the performance is better when employing prioritized progressive search.

B. Comparison on synthetic datasets

The comparative methods include the RANSAC-based methods EPnP [21], P3P [20], 2-Entity [28] and globally optimal method LMI [29]. We use the OpenCV [38] implementation of EPnP and P3P. For LMI, we modify their open source code in MATLAB following the paper, since only code for 3D-3D registration is released. In addition, we control the evaluation data having rotation angle less than 60° and add it as the constraint of LMI, as suggested in [29]. The 2-Entity RANSAC is implemented in MATLAB and we select the mixed sampling strategy which utilize both points and lines for pose estimation. All methods are followed by nonlinear refinement on the identified consesus set. We still use the synthetic dataset as in the ablation study.



(a)

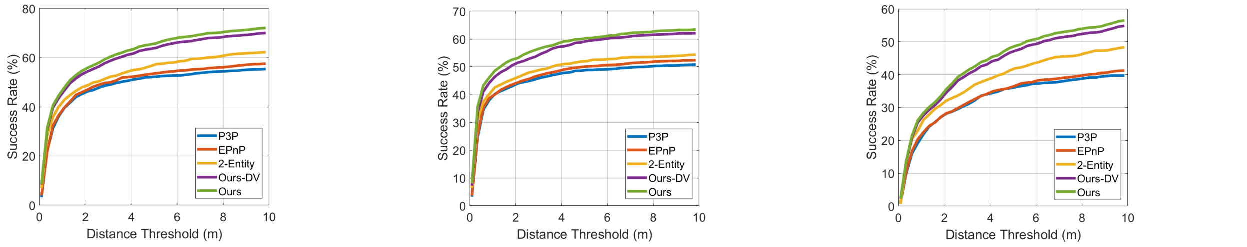
(b)

Fig. 6: (a) The number of inliers in the estimated maximal consensus set w.r.t increasing outliers of successful estimation. (b) The number of inliers in the estimated maximal consensus set for 100 runs when the outlier rate is 80%.

Efficiency of globally optimal methods. We first compare the efficiency between the proposed method and the LMI. We evaluate the computational cost with respect to the number of feature correspondences and the percentage of outliers. The result is shown in Fig. 4, the computational cost of LMI is significantly higher than the proposed methods both for increasing number of correspondences, and the percentage of outliers. The growing gap may also indicate that the complexity of LMI is higher than ours.

Deterministic convergence. The vital difference between RANSAC and globally optimal method is the convergence. We compare the number of inliers in the estimated maximal consensus set with respect to increasing outliers when the final pose estimation is successful. The result is shown in Fig. 6, which indicates that the proposed solution achieves deterministic perfect CCI, while RANSAC gives conservative estimations with less inliers and LMI finds optimistic estimations by incorrectly regarding outliers as inliers. In addition, both RANSAC and LMI fail when the outlier rate is 90%. The results for all 100 runs when the outlier rate is 80% are also shown in Fig. 6. We can see that the proposed algorithm deterministically finds the globally optimal consensus, while RANSAC achieves global optimality probabilistically.

Robustness and accuracy. We finally show the performance of all methods on the synthetic data, including accuracy, precision and recall to measure the CCI, with respect to percentage of outliers ranging from 60% to 90%. Note that we only evaluate the accuracy for successful trials, since result on incorrectly identified consensus set can lead to very large error, disturbing the accuracy. The result in Tab. I first confirms that CCI is highly related to the accuracy, validating the feasibility of maximizing consensus set. RANSAC gives consistent conservative estimations, as the precision remains at a higher level compared with the recall. For LMI, the estimation is prone to regard the outliers as inliers, thus the



(a)

(b)

(c)

Fig. 7: Success rate with respect to threshold on the whole three sessions 0827 (left), 0828 (center) and 0129 (right).

TABLE I: Accuracy and CCI comparison.

Outlier	Method	P3P	EPnP	2-Entity	LMI	Ours-DV	Ours
60%	△T(m)	0.0010	0.0009	0.0008	0.0128	**0.0005**	0.0006
	△R(°)	0.0196	0.0170	0.0059	0.0083	**0.0019**	0.0020
	Precision	**1.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**
	Recall	0.99	0.99	**1.00**	0.98	**1.00**	**1.00**
	Success%	**100**	**100**	**100**	65	**100**	**100**
70%	△T(m)	0.0013	-	0.0011	0.0209	**0.0005**	0.0006
	△R(°)	0.0213	-	0.0211	0.1059	**0.0017**	0.0028
	Precision	**1.00**	0	**1.00**	0.93	**1.00**	**1.00**
	Recall	0.98	0	0.99	0.93	**1.00**	**1.00**
	Success%	**100**	0	**100**	54	**100**	**100**
80%	△T(m)	0.0017	-	0.0017	0.0246	0.0007	**0.0006**
	△R(°)	0.0267	-	0.0257	0.4778	0.0050	**0.0032**
	Precision	**1.00**	0	**1.00**	0.46	**1.00**	**1.00**
	Recall	0.49	0	0.93	0.58	**1.00**	**1.00**
	Success%	52	0	96	37	**100**	**100**
90%	△T(m)	-	-	0.0027	-	**0.0007**	**0.0007**
	△R(°)	-	-	0.0411	-	0.0073	**0.0043**
	Precision	0	0	**1.00**	0.27	**1.00**	**1.00**
	Recall	0	0	0.70	0.35	**1.00**	**1.00**
	Success%	0	0	86	0	**100**	**100**

[1] The accuracy is evaluated for successful trials, the precision and recall of CCI are for all test trails.
[2] Ours-DV denotes the proposed method with dimension-wise voting.

TABLE II: Performance on selected cases in real world.

	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$
	01	9/18	0/0	02	15/39	0/0
Method	$\triangle T$	$\triangle R$	Inliers[1]	$\triangle T$	$\triangle R$	Inliers[1]
	(m)	(°)	$\vert\zeta^*\vert/\vert\zeta\vert$	(m)	(°)	$\vert\zeta^*\vert/\vert\zeta\vert$
EPnP	0.9938	0.8025	7/12	0.9026	1.3255	11/21
P3P	0.8187	0.6302	7/11	1.9751	0.5977	10/20
2-Entity	0.6683	0.4351	8/10	0.5703	0.3378	12/21
LMI	0.1630	0.1951	9/13	0.2832	0.2155	14/19
Ours-DV	0.1207	0.1321	9/ 9	0.1803	0.1550	14/14
Ours	0.1207	0.1321	9/ 9	0.1753	0.1334	15/15
	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$
	03	21/65	0/2	04	23/48	7/15
EPnP	0.4506	0.9741	10/29	0.5504	0.7823	19/28
P3P	0.3213	0.8807	13/27	0.3678	0.4066	19/27
2-Entity	0.3138	0.4603	15/27	0.1405	0.2055	27/33
LMI	0.2998	0.3786	19/44	0.2834	0.1769	22/28
Ours-DV	0.1407	0.1743	21/23	0.0309	0.1607	28/29
Ours	0.1382	0.1707	21/23	0.0253	0.1509	30/30
	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$	ExpID	$\vert\zeta_P\vert/N_P$	$\vert\zeta_L\vert/N_L$
	05	21/38	8/13	06	96/134	3/4
EPnP	1.0876	0.8111	13/25	0.2705	0.5202	93/112
P3P	1.0876	0.8111	13/25	0.1682	0.5243	90/98
2-Entity	0.1732	0.2687	27/29	0.1163	0.4623	95/108
LMI	0.7641	0.6394	16/28	0.0891	0.2812	96/102
Ours-DV	0.1671	0.1072	29/29	0.0861	0.2791	99/99
Ours	0.1671	0.1072	29/29	0.0861	0.2791	99/99

[1] $|\zeta|$ denotes the number of identified inliers, while $|\zeta^*|$ the true inliers.

recall is higher compared with precision. Considering that LMI, P3P and EPnP are designed for general visual localization, the better performance achieved by 2-Entity and the proposed method, designed for visual inertial localization, is reasonable. But we can still summarize that superior result can be found by specialized globally optimal method.

C. Comparison on visual inertial localization

Finally, we evaluate all the methods on a real world cross-session visual inertial localization task. The dataset employed is YQ-dataset [39]. In the dataset, there are three sessions collected in summer 2017, denoted as 2017-0823, 2017-0827 and 2017-0828, and one session in winter 2018 after snow denoted as 2018-0129. The 3D map is built with 2017-0823 session and the other three sessions are used to evaluate the localization performance, indicating the changing environment. The details to obtain the 3D-2D point and line correspondences can be found in Appendix. For evaluation, we compute the ground truth relative pose between the query camera and the map by aligning the synchronized LiDAR scans. For the pitch and roll angle, we use the estimation of visual inertial odometry [40].

Selected cases performance. We first select several typical examples for evaluation as in [29] and the results are shown in Tab. II. The Exp01, Exp02 and Exp03 are cases with pure point features where Exp03 has lines as disturbance and the outlier rate in these three cases are all more than 50%. The RANSAC-based methods perform poorly compared with the global optimization methods. One thing to note is that in real world dataset, dimension-wise voting brings slight performance drop, but still achieves superior performance against comparative methods. Also note that in Exp03, the proposed method gives optimistic results by regarding 2 outliers as inliers, which may be caused by unknown noise bound thus inappropriate threshold in real world data. In Exp04, Exp05 and Exp06, the utilization of good line features promotes the performance of point line methods obviously (2-Entity and ours). Overall, the results still confirm the conclusions in simulation.

Full dataset performance. Finally, we arrive at the success rate on the whole three sessions as shown in Fig. 7. As LMI is too slow to finish all the dataset, here we only show the result of ours and RANSAC methods. We first see that the proposed globally optimal methods consistently outperform the RANSAC methods on all three sessions. The other fact is that progressive prioritized voting brings the best accuracy over the one with dimension-wise voting, because of the consideration on extremely low number of inliers.

VI. Conclusions

In this paper, we propose a robust solver designed for visual inertial localization problem, achieving global optimization of the consensus maximization problem with deterministic convergence, even when the percentage of outliers is very high, say 90%. The key step in our solver is the derivation of *translation invariant measurements* for both points and lines, thus decoupling the problem into two smaller subproblems. Then we propose 1D BnB and prioritized progressive voting to find globally optimal rotation and translation respectively, accelerating the search efficiency. The effectiveness of the proposed method is validated on both synthetic and real world dataset.

Appendix I
Derivation of TIMs

With the aid of inertial measurements, the pitch and roll angle between the current query camera frame and the gravity-aligned world reference frame are globally observable, such that the rotation estimation of the query camera with respect to the world can be formulated as

$$\begin{aligned} R_{\mathcal{WC}} &= R_z(\alpha)R_y(\check{\beta})R_x(\check{\gamma}) \\ &= \begin{bmatrix} c\alpha & -s\alpha & 0 \\ s\alpha & c\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c\check{\beta} & 0 & s\check{\beta} \\ 0 & 1 & 0 \\ -s\check{\beta} & 0 & c\check{\beta} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & c\check{\gamma} & -s\check{\gamma} \\ 0 & s\check{\gamma} & c\check{\gamma} \end{bmatrix} \\ &\triangleq \begin{bmatrix} a_{11}c\alpha + b_{11}s\alpha & a_{12}c\alpha + b_{12}s\alpha & b_{13}s\alpha \\ a_{21}c\alpha + b_{21}s\alpha & a_{22}c\alpha + b_{22}s\alpha & a_{23}c\alpha \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \end{aligned} \tag{27}$$

where $\check{\beta}$ and $\check{\gamma}$ denote the observed pitch and roll angle provided by inertial measurements, α denotes the yaw angle to be estimated, $\sin\alpha \triangleq s\alpha$, $\cos\alpha \triangleq c\alpha$. Therefore, the rotation matrix is only determined by the estimation of yaw, which is the same in R, as $R = R_{\mathcal{WC}}^T$. Thus the degrees of freedom (DoF) of the rotation matrix estimation can be reduced to 1 with the aid of inertial measurements, that is

$$R = R(\alpha) = \begin{bmatrix} a_{11}c\alpha + b_{11}s\alpha & a_{21}c\alpha + b_{21}s\alpha & a_{31} \\ a_{12}c\alpha + b_{12}s\alpha & a_{22}c\alpha + b_{22}s\alpha & a_{32} \\ b_{13}s\alpha & a_{23}c\alpha & a_{33} \end{bmatrix} \tag{28}$$

A. Derivation of point-TIM

The collinearity of each 2D-3D point features is utilized to derive the point-TIM as shown in Fig. 8. Mathematically, given an image key point u_i, we have an un-normalized direction vector from the camera center as

$$\tilde{u}_i \triangleq \begin{pmatrix} \tilde{u}_{i,x} \\ \tilde{u}_{i,y} \\ 1 \end{pmatrix} = K^{-1} \begin{pmatrix} u_i \\ 1 \end{pmatrix} \tag{29}$$

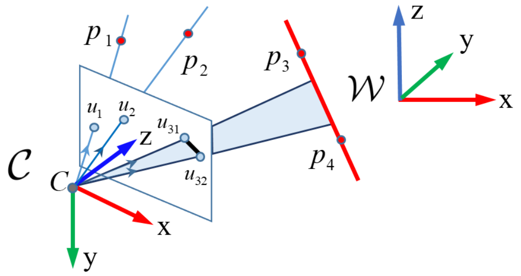


Fig. 8: The illustration of 2D-3D point and line features.

According to the projection geometry, the optical center of camera frame $C = \mathbf{0}_{3\times1}$, the 2D point $\tilde{u}_1$ and the corresponding 3D point p_1 lie on the same line, which is denoted as $\{C, \tilde{u}_1, Rp_1 + t\}_L$. By solving the line equation from the first two points and substituting the third point into the equation, we have

$$\frac{R_1 p_1 + t_x}{\tilde{u}_{1,x}} = \frac{R_2 p_1 + t_y}{\tilde{u}_{1,y}} = R_3 p_1 + t_z \tag{30}$$

where $R \triangleq (R_1^T, R_2^T, R_3^T)^T$ and $t \triangleq (t_x, t_y, t_z)^T$. Based on (30), we have two constraints from a correspondence as

$$\tilde{u}_{1,x}(R_2 p_1 + t_y) - u_{1,y}(R_1 p_1 + t_x) = 0 \tag{31}$$

$$\tilde{u}_{1,x}(R_3 p_1 + t_z) - (R_1 p_1 + t_x) = 0 \tag{32}$$

Naturally, given another correspondence u_2 and p_2, according to $\{C, \tilde{u}_2, Rp_2 + t\}_L$

$$\frac{R_1 p_2 + t_x}{\tilde{u}_{2,x}} = \frac{R_2 p_2 + t_y}{\tilde{u}_{2,y}} = R_3 p_2 + t_z \tag{33}$$

Then we can have two more constraints as

$$\tilde{u}_{2,x}(R_2 p_2 + t_y) - u_{2,y}(R_1 p_2 + t_x) = 0 \tag{34}$$

$$\tilde{u}_{2,x}(R_3 p_2 + t_z) - (R_1 p_2 + t_x) = 0 \tag{35}$$

Combining (31) - (32), t_y and t_z can be eliminated, then substituted into (34) - (35), t_x can also be eliminated, resulting in an constraint only relating to R. Recall (28), by reorganizing the coefficients, we have the point-TIM as

$$d_p(\alpha) = d_{p,1}\sin\alpha + d_{p,2}\cos\alpha + d_{p,3} \tag{36}$$

B. Derivation of line-TIM

Each line feature correspondence can be represented by a pair of start point and end point of the line segment as shown in Fig. 8. According to the projection geometry, the optical center of the camera, the 2D line segment (u_{31}, u_{32}) and the 3D line (p_3, p_4) lie on the same plane. Then the four points C, u_{31}, u_{32} and p_3 are coplanar, denoted as $\{C, u_{31}, u_{32}, p_3\}_P$. Similarly, $\{C, u_{31}, u_{32}, p_4\}_P$ also holds. By solving the plane equation from the first three points and substituting the fourth point into it, we have:

$$(\tilde{u}_{31} \times \tilde{u}_{32})^T (Rp_3 + t) = 0 \tag{37}$$

That is

$$\begin{aligned} &(u_{31,y} - u_{32,y})(R_1 p_3 + t_x) - (u_{31,x} - u_{32,x})(R_2 p_3 + t_y) \\ &+ (u_{31,x}u_{32,y} - u_{32,x}u_{31,y})(R_3 p_3 + t_z) = 0 \end{aligned} \tag{38}$$

Similarly, for $\{C, u_{31}, u_{32}, p_4\}_P$, we have:

$$(\tilde{u}_{31} \times \tilde{u}_{32})^T (Rp_4 + t) = 0 \tag{39}$$

That is

$$\begin{aligned}(u_{31,y} - u_{32,y})(R_1 p_4 + t_x) - (u_{31,x} - u_{32,x})(R_2 p_4 + t_y) \\ + (u_{31,x} u_{32,y} - u_{32,x} u_{31,y})(R_3 p_4 + t_z) = 0\end{aligned} \tag{40}$$

With (38)-(40), the t can be eliminated resulting in

$$\begin{aligned}&[(u_{31,y} - u_{32,y})R_1 - (u_{31,x} - u_{32,x})R_2 \\ &+ (u_{31,x} u_{32,y} - u_{32,x} u_{31,y})R_3](p_3 - p_4) = 0\end{aligned} \tag{41}$$

Recall (28), (41) can be reorganized to line-TIM as

$$d_l(\alpha) = d_{l,1} \sin\alpha + d_{l,2} \cos\alpha + d_{l,3} \tag{42}$$

C. Derivation of TIMs' lower bound

Recall (36) and (42), as the forms of point-TIM and line-TIM are the same, we denote them as $d(\alpha)$. That is

$$\begin{aligned}d(\alpha) &= d_1 \sin\alpha + d_2 \cos\alpha + d_3 \\ &= \sqrt{d_1^2 + d_2^2}(\sin\alpha \cos a_2 + \cos\alpha \sin a_2) + d_3 \\ &= a_1 \sin(\alpha + a_2) + d_3\end{aligned} \tag{43}$$

where $a_1 = \sqrt{d_1^2 + d_2^2}$, $\sin a_2 = \frac{d_2}{a_1}$, $\cos a_2 = \frac{d_1}{a_1}$.

Then the lower bound of $|d(\alpha)|$ on $\mathbb{A}$, denoted as $\underline{d}(\mathbb{A})$, is derived as

$$\underline{d}(\mathbb{A}) = \min |a_1 \sin(\alpha + a_2) + d_3| \tag{44}$$

Appendix II
Derivation of Translation Bound

After the rotation estimation, we get the optimal yaw angle $\hat{\alpha}$. As shown in Fig. 8, according to $\{C, \tilde{u}_1, R(\hat{\alpha})p_1 + t\}_L$, we have

$$\tilde{u}_1 \times (R(\hat{\alpha})p_1 + t) = 0 \tag{45}$$

which is equal to

$$\tilde{u}_{1\times}(R(\hat{\alpha})p_1 + t) = 0 \tag{46}$$

where $a_\times$ denotes the symmetric matrix of vector a. Then (46) can be written as

$$\begin{bmatrix} 0 & -1 & \tilde{u}_{1,y} \\ 1 & 0 & -\tilde{u}_{1,x} \\ -\tilde{u}_{1,y} & \tilde{u}_{1,x} & 0 \end{bmatrix} \begin{bmatrix} t_x + h_1 \\ t_y + h_2 \\ t_z + h_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{47}$$

where $R(\hat{\alpha})p_1 \triangleq (h_1, h_2, h_3)^T$. Then two equations of translation can be derived as

$$-t_y - h_2 + \tilde{u}_{1,y}(t_z + h_3) = 0 \tag{48}$$

$$t_x + h_1 - \tilde{u}_{1,x}(t_z + h_3) = 0 \tag{49}$$

Similarly, with another point correspondence $\{C, \tilde{u}_2, R(\hat{\alpha})p_2 + t\}_L$, we have

$$\begin{bmatrix} 0 & -1 & \tilde{u}_{2,y} \\ 1 & 0 & -\tilde{u}_{2,x} \\ -\tilde{u}_{2,y} & \tilde{u}_{2,x} & 0 \end{bmatrix} \begin{bmatrix} t_x + h_4 \\ t_y + h_5 \\ t_z + h_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{50}$$

where $R(\hat{\alpha})p_2 \triangleq (h_4, h_5, h_6)^T$. Then we have another two equations as

$$-t_y - h_5 + \tilde{u}_{2,y}(t_z + h_6) = 0 \tag{51}$$

$$t_x + h_4 - \tilde{u}_{2,x}(t_z + h_6) = 0 \tag{52}$$

Combining (48)-(49) and (51)-(52), the translation can be solved as

$$t_x = \frac{\tilde{u}_{1,x}}{\tilde{u}_{1,y} - \tilde{u}_{2,y}}(\tilde{u}_{2,y}(h_6 - h_3) + h_2 - h_5) - h_1 \tag{53}$$

$$t_y = \frac{\tilde{u}_{1,y}}{\tilde{u}_{1,y} - \tilde{u}_{2,y}}(\tilde{u}_{2,y}(h_6 - h_3) + h_2 - h_5) + h_2 \tag{54}$$

$$t_z = \frac{1}{\tilde{u}_{1,y} - \tilde{u}_{2,y}}(h_2 - h_5 - \tilde{u}_{1,y} h_3 + \tilde{u}_{2,y} h_6) \tag{55}$$

In addition, the translation can also be solved with one point and one line correspondence. According to (37)

$$(\tilde{u}_{31} \times \tilde{u}_{32})^T (R(\hat{\alpha})p_3 + t) = 0 \tag{56}$$

we have

$$\tilde{u}_{31} \times \tilde{u}_{32} = \begin{bmatrix} \tilde{u}_{1,y} - \tilde{u}_{2,y} \\ \tilde{u}_{1,x} + \tilde{u}_{2,x} \\ \tilde{u}_{1,x}\tilde{u}_{2,y} - \tilde{u}_{2,x}\tilde{u}_{1,y} \end{bmatrix} \triangleq \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{57}$$

Then (37) can be written as

$$\begin{bmatrix} n_1 & n_2 & n_3 \end{bmatrix} \begin{bmatrix} t_x + m_1 \\ t_y + m_2 \\ t_z + m_3 \end{bmatrix} = 0 \tag{58}$$

where $R(\hat{\alpha})p_3 + t \triangleq (m_1, m_2, m_3)^T$. Similarly, with (39), we have

$$\begin{bmatrix} n_1 & n_2 & n_3 \end{bmatrix} \begin{bmatrix} t_x + m_4 \\ t_y + m_5 \\ t_z + m_6 \end{bmatrix} = 0 \tag{59}$$

where $R(\hat{\alpha})p_4 + t \triangleq (m_4, m_5, m_6)^T$. Thus, combining (48)-(49) and (58)-(59), the translation can be solved as

$$\begin{aligned}t_z =& (-n_1\tilde{u}_{1,x}h_3 + n_1 h_1 - n_1 m_1 - n_2\tilde{u}_{1,y}h_3 - n_2 h_2 \\ &- n_2 m_2 - n_3 m_3)/(n_1\tilde{u}_{1,x} + n_2\tilde{u}_{1,y} + n_3)\end{aligned} \tag{60}$$

$$t_y = \tilde{u}_{1,y}(t_z + h_3) + h_2 \tag{61}$$

$$t_x = \tilde{u}_{1,x}(t_z + h_3) - h_1 \tag{62}$$

Recall (1), there is *unknown but bounded* [41] noise on the detected image features, such that $|e_i| < n_i$, and we have

$$\underline{u}_i = u_i - n_i, \overline{u}_i = u_i + n_i \tag{63}$$

With this feature bound of u_i, the un-normalized direction vector $\tilde{u}_i$ can also be bounded after linear transformations. Then the bound of the derived translation $[\ \underline{t}_{ij}, \overline{t}_{ij}\ ]$ can be computed with the following relaxation [34] [35]:

$$\begin{aligned}&f = ab,\ \underline{a} \le a \le \overline{a},\ \underline{b} \le b \le \overline{b}, \\ &f \ge \max(\underline{a}b + \underline{b}a - \underline{a}\underline{b}, \overline{a}b + \overline{b}a - \overline{a}\overline{b}) \\ &f \le \min(\overline{a}b + \underline{b}a - \overline{a}\underline{b}, \underline{a}b + \overline{b}a - \underline{a}\overline{b})\end{aligned} \tag{64}$$

Appendix III
Real World Experiment Details

The dataset employed in real world cross-session visual inertial localization task is YQ-dataset [39]. In the dataset, there are three sessions collected at summer 2017 in three days, denoted as 2017-0823, 2017-0827 and 2017-0828, and one session collected in winter 2018 after snow, denoted as 2018-0129. The 3D map is built with 2017-0823 session and the 3D-2D point feature correspondences are obtained by running visual inertial SLAM [40]. For evaluation, we compute the ground truth of the relative pose between the query camera and the map by aligning the synchronized LiDAR scans. For the pitch and roll angle, we use the estimation generated by visual inertial odometry [40]. To get the 3D-2D feature matches between the query image and the map, we exploited the following steps:

- Obtain the camera poses and the 3D-2D point matches in the map using visual inertial SLAM software [40].
- Run Line3D++ algorithm [42] to get the 3D-2D line matches in the map.
- For the query session, we get the 3D-2D points/lines match based on the descriptors of LibVISO2 [13] and LBD [43].

References

[1] M. Li and A. I. Mourikis, "High-precision, consistent ekf-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[2] 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.

[3] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[4] S. Choi, T. Kim, and W. Yu, "Performance evaluation of ransac family," *Journal of Computer Vision*, vol. 24, no. 3, pp. 271–300, 1997.

[5] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim, "Robust regression methods for computer vision: A review," *International journal of computer vision*, vol. 6, no. 1, pp. 59–70, 1991.

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

[7] M. Brown, D. Windridge, and J.-Y. Guillemaut, "Globally optimal 2d-3d registration from points or lines without correspondences," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2111–2119, 2015.

[8] D. Campbell, L. Petersson, L. Kneip, and H. Li, "Globally-optimal inlier set maximisation for simultaneous camera pose and feature correspondence," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–10, 2017.

[9] C. Olsson, F. Kahl, and M. Oskarsson, "Branch-and-bound methods for euclidean registration problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 783–794, 2008.

[10] H. Li, "Consensus set maximization with guaranteed global optimality for robust geometry estimation," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1074–1080, IEEE, 2009.

[11] T.-J. Chin, Y. Heng Kee, A. Eriksson, and F. Neumann, "Guaranteed outlier removal with mixed integer linear programs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5858–5866, 2016.

[12] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, Ieee, 2004.

[13] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium (IV)*, 2011.

[14] P. Furgale and T. D. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010.

[15] L. Tang, Y. Wang, X. Ding, H. Yin, R. Xiong, and S. Huang, "Topological local-metric framework for mobile robots navigation: a long term perspective," *Autonomous Robots*, pp. 1–15, 2018.

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

[17] 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*, vol. 1, 2015.

[18] T. Schneider, M. 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*, vol. 3, no. 3, pp. 1418–1425, 2018.

[19] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[20] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[22] P. Wang, G. Xu, Z. Wang, and Y. Cheng, "An efficient solution to the perspective-three-point pose problem," *Computer Vision and Image Understanding*, vol. 166, pp. 81–87, 2018.

[23] M. Dhome, M. Richetin, J.-T. Lapreste, and G. Rives, "Determination of the attitude of 3d objects from a single perspective view," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 1265–1278, 1989.

[24] H. H. Chen, "Pose determination from line-to-plane correspondences: Existence condition and closed-form solutions," in *Computer Vision, 1990. Proceedings, Third International Conference on*, pp. 374–378, IEEE, 1990.

[25] S. Ramalingam, S. Bouaziz, and P. Sturm, "Pose estimation using both points and lines for geo-localization," in *ICRA 2011-IEEE International Conference on Robotics and Automation*, pp. 4716–4723, IEEE Computer Society, 2011.

[26] L. Kneip, M. Chli, and R. Y. Siegwart, "Robust real-time visual odometry with a single camera and an imu," in *Proceedings of the British Machine Vision Conference 2011*, British Machine Vision Association, 2011.

[27] Z. Kukelova, M. Bujnak, and T. Pajdla, "Closed-form solutions to minimal absolute pose problems with known vertical direction," in *Asian Conference on Computer Vision*, pp. 216–229, Springer, 2010.

[28] Y. Jiao, Y. Wang, B. Fu, X. Ding, Q. Tan, L. Chen, and R. Xiong, "2-entity ransac for robust visual localization in changing environment," *arXiv preprint arXiv:1903.03967*, 2019.

[29] P. Speciale, D. Pani Paudel, M. R. Oswald, T. Kroeger, L. Van Gool, and M. Pollefeys, "Consensus maximization with linear matrix inequality constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4941–4949, 2017.

[30] T. M. Breuel, "Implementation techniques for geometric branch-and-bound matching methods," *Computer Vision and Image Understanding*, vol. 90, no. 3, pp. 258–294, 2003.

[31] J. Yang, H. Li, D. Campbell, and Y. Jia, "Go-icp: A globally optimal solution to 3d icp point-set registration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2241–2254, 2015.

[32] Y. Liu, C. Wang, Z. Song, and M. Wang, "Efficient global point cloud registration by matching rotation invariant features through translation search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 448–463, 2018.

[33] H. Yang and L. Carlone, "A polynomial-time solution for robust registration with extreme outlier rates," *arXiv preprint arXiv:1903.08588*, 2019.

[34] G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems," *Mathematical programming*, vol. 10, no. 1, pp. 147–175, 1976.

[35] H. D. Sherali and A. Alameddine, “A new reformulation-linearization technique for bilinear programming problems,” *Journal of Global optimization*, vol. 2, no. 4, pp. 379–410, 1992.

[36] P. Miraldo, T. Dias, and S. Ramalingam, “A minimal closed-form solution for multi-perspective pose estimation using points and lines,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 474–490, 2018.

[37] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, “Robust visual inertial odometry using a direct ekf-based approach,” in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 298–304, IEEE.

[38] https://opencv.org/.

[39] X. Ding, Y. Wang, D. Li, L. Tang, H. Yin, and R. Xiong, “Laser map aided visual inertial localization in changing environment,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4794–4801, IEEE, 2018.

[40] 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.

[41] M. Milanese, “Estimation and prediction in the presence of unknown but bounded uncertainty: a survey,” in *Robustness in Identification and Control*, pp. 3–24, Springer, 1989.

[42] M. Hofer, M. Maurer, and H. Bischof, “Efficient 3d scene abstraction using line segments,” *Computer vision and image understanding*, vol. 157, pp. 167–178, 2017.

[43] L. Zhang and R. Koch, “An efficient and robust line segment matching approach based on lbd descriptor and pairwise geometric consistency,” *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794–805, 2013.