GVIL: Tightly Coupled GNSS-Visual-Inertial-Lidar for Position Estimation in Challenging Environments

Yanfang Shi; Baowang Lian
School of Electronics And Information
Northwestern Polytechnical University
Shaanxi, China
syf1819285@mail.nwpu.edu.cn
bwlian@nwpu.edu.cn

Yonghong Zeng; Ernest Kurniawan; Yugang Ma
Institute for Infocomm Research
Agency for Science, Technology and Research
Singapore, Singapore
yhzeng@i2r.a-star.edu.sg
ekurniawan@i2r.a-star.edu.sg
mayg@i2r.a-star.edu.sg

***Abstract*—The fusion framework based on Light Detection and Ranging (LiDAR) and vision is susceptible to environmental constraints. Therefore, they are unable to adapt to complex and challenging environments. In response to this, this paper proposes a tightly coupled framework based on Global Navigation Satellite System (GNSS) data, which tightly couples GNSS, inertial, LiDAR and visual data. The experimental results demonstrate that, using publicly available challenging environment datasets, the proposed fusion system can achieve a 52.68% improvement in positioning accuracy compared to some existing fusion algorithms.**

***Index Terms*—Fusion, GNSS, Inertial, LiDAR, Tightly-coupled, Vision.**

I. Introduction

RELATIVE positioning systems based on LiDAR and vision are affected by environmental conditions, and inertial sensors are prone to estimation accuracy degradation due to error accumulation, rendering them inadequate for complex scenarios [1], [2]. Some researchers began to try combining GNSS absolute positioning and relative methods to improve position estimation accuracy in harsh environments. Broadly, it can be categorized into loosely coupled and tightly coupled methods.

1) Loosely-coupled Fusion Involving GNSS: Reference [3] presents a new comparison-based algorithm of pseudorange in order to improve GNSS/IMU (Inertial Measurement Unit) integrated positioning. Reference [4] presents one GNSS/INS/Vision solution, including vision-based detection and tracking and position estimation and optimization. Reference [5] proposes a loosely-coupled pipeline that fuses visual, inertial, LiDAR, and GNSS data by Unscented Kalman Filter (UKF) algorithm to produce a smooth and consistent trajectory in different environments. However, in the above methods, GNSS is susceptible to electromagnetic interference signals and environmental obstructions that hinder achieving precise positioning.

2) Tightly-coupled Fusion Involving GNSS: In reference [6], the Extended Kalman Filter (EKF) method is used to fuse GNSS raw data with other sensor data. [7] and [8] adopt the optimization method to integrate GNSS raw data with other data. However, the robustness of the fusion framework using filtering algorithms to tightly couple GNSS raw data still cannot meet the requirements in challenging environments.

II. Our Contribution

- This paper proposes a tightly coupled framework based on nonlinear optimization, integrating GNSS raw data, inertial data, LiDAR data, and visual data. The framework initially derives spatial parameters for other sensors relative to the carrier's center and then performs position calculations based on this foundation, saving computation time and improving efficiency.
- The proposed tightly coupled framework utilizes GNSS pseudorange to construct GNSS optimization factors. Even with only one available satellite data, it can still participate in the computation. This enables the fusion system to fully utilize available GNSS data in environments where satellite signals are unfavorable, making it more adaptable to challenging environments.
- This paper constructs visual optimization factors using the reprojection principle and constructs LiDAR opti-

This research is supported in part by National Natural Science Foundation of China under Grant 62173276, 62171735, 62101458, 62001392, 61803310 and 61801394, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022GY-097,2021JQ-122 and 2021JQ-693, and in part by China Postdoctoral Science Foundation under Grant 2020M673482 and 2020M673485. This research is also supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme (FCP-NTU-RG-2022-021, FCP-ASTAR-TG-2022-003). The author would like to express gratitude for the support and sponsorship provided by the China Scholarship Council (Liu Jin Xuan [2021]70). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

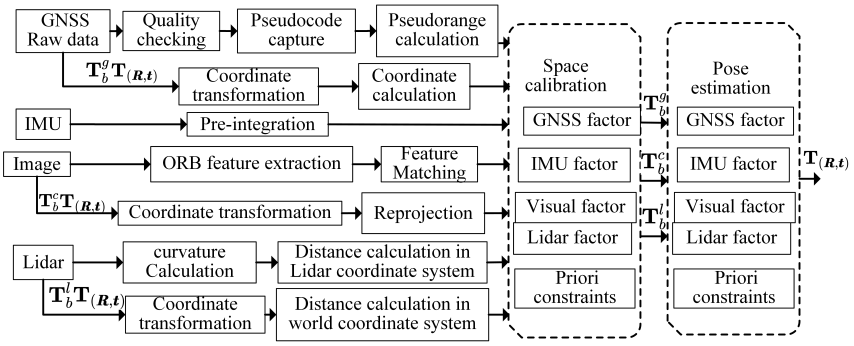


Fig. 1. Schematic diagram of the overall workflow of the proposed fusion algorithm

mization factors based on the principle that the point-to-line (or surface) distances are equal in different coordinate systems. Therefore, it enables the full utilization of visual and LiDAR data.

- Finally, using publicly available datasets, simulation results demonstrate that the proposed fusion framework can improve positioning accuracy by 52.68% compared to existing open-source programs. Therefore, the proposed fusion system exhibits greater robustness in challenging environments than existing algorithms.

III. Problem Formulation

The specific workflow diagram of the proposed fusion algorithm is shown in Fig. 1. The proposed fusion algorithm consists of two steps. The first step determines $\mathbf{T}_b^c$, $\mathbf{T}_b^l$, $\mathbf{T}_b^g$, which are the spatial parameters between the inertial data and visual (camera) data, LiDAR data, and GNSS data, respectively. The second step, building upon the first, calculates the transformation matrix between positions at different time points $\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}$, whose composition is defined by Eq. (1).

$$\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \tag{1}$$

The matrix $\boldsymbol{R}$ represents the rotation matrix of dimensions 3×3, indicating the carrier's orientation, while the vector $\boldsymbol{t}$ with dimensions 3×1 represents the carrier's translation.

Inspired by [9] and [10], based on the Maximum A Posteriori (MAP) estimation, we can transform the location estimation problem of the fusion system into a probability problem. Namely, under the assumption of the actual state, we aim to maximize the probability of the currently measured data. Assuming the noise follows a Gaussian distribution, through Bayesian theory and probability formulas, the location estimation can be transformed into a minimization problem.

Then we decompose the optimization problem into individual factors related to states and measurements [11]. Therefore, the core of the proposed fusion algorithm lies in establishing factors for each type of sensor data.

IV. The Optimization Factors

The detailed explanations for the celestial coordinate systems can be found in [10]. The optimization factor $\mathbf{f}_b(\mathbf{z}_b)$ for the inertial sensor has been derived in [9], hence it is not elaborated in the following subsections.

A. Vision Factor

This paper first extracts Oriented FAST and Rotated BRIEF (ORB) feature points [12] from visual images. Only when the landmark point is projected to pixels corresponding to ORB feature points is it selected to participate in the reprojection process and feature-matching algorithm.

Firstly, we multiply the world coordinate position of the carrier at the previous moment $\boldsymbol{P}_{b,k-1}^w$ by the spatial parameters $\mathbf{T}_b^c$ between the inertial sensor and the camera to obtain the position of the camera in the world coordinate system at the previous time $\boldsymbol{P}_{c,k-1}^w$. Through the projection algorithm, we can get the pixel coordinates of this landmark on the previous picture, as shown in Eq. (2).

$$\boldsymbol{y}_{k-1} = \boldsymbol{\pi}_c(\boldsymbol{P}_{c,k-1}^w) = \boldsymbol{\pi}_c(\mathbf{T}_b^c\boldsymbol{P}_{b,k-1}^w) \tag{2}$$

where, $\boldsymbol{\pi}_c$ is the projection function of the camera. Then, we adopt the feature points matching method [12] to find the pixel coordinates $\boldsymbol{y}_k$ of the current moment corresponding to the pixel $\boldsymbol{y}_{k-1}$ of the previous moment, which is taken as the actual measurement.

Similarly, we multiply the world coordinate position of the carrier at the previous time $\boldsymbol{P}_{b,k-1}^w$ by rotation and translation of the carrier between the two moments $\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}$. In order to obtain the world coordinate position of the carrier at the current time $\boldsymbol{P}_{b,k}^w$. In the same way, we can also get the pixel coordinates of the camera on the current picture, as shown in Eq. (3), which is then taken as the predicted value.

$$\widetilde{\boldsymbol{y}}_k = \boldsymbol{\pi}_c(\boldsymbol{P}_{c,k}^w) = \boldsymbol{\pi}_c(\mathbf{T}_b^c\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}_{b,k-1}^w) \tag{3}$$

Therefore, the vision optimization factor is the error between the predicted and measured pixel values, as shown in Eq. (4).

$$\begin{aligned} \mathbf{f}_v(\mathbf{z}_v) &= \sum_{i=1}^{m}(\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i) \\ &= \sum_{i=1}^{m}\{\boldsymbol{\pi}_c(\mathbf{T}_b^c\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}_{b,i-1}^w) - \mathrm{Match}(\boldsymbol{\pi}_c(\mathbf{T}_b^c\boldsymbol{P}_{b,i-1}^w))\} \end{aligned} \tag{4}$$

m represents the number of landmark points.

B. LiDAR Factor

We extract corner and plane points of the points cloud [13]. Subsequently, we employ the distance calculation formula to obtain the point-to-line or point-to-plane distance.

We know that the distance should remain the same in different coordinate systems. On the one hand, we can directly obtain the corresponding points cloud coordinates through measurements and obtain the distance values in the LiDAR coordinate system, which serves as the actual measurement values. We denote the distance of point-to-line (point-to-plane) as $d_{\varepsilon m}(d_{\zeta_m})$.

On the other hand, multiplying $\boldsymbol{P}^w_{b,k-1}$ by $\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}$, we can obtain the carrier's position at time k in the world coordinate system, as shown in Eq. (5).

$$\boldsymbol{P}^w_{b,k} = \mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}^w_{b,k-1} \tag{5}$$

Subsequently, multiplying $\boldsymbol{P}^w_{b,k}$ by $\mathbf{T}^l_b$, we can obtain the LiDAR sensor's position at time k in the world coordinate system $\boldsymbol{P}^w_{l,k}$, as shown in Eq. (6).

$$\boldsymbol{P}^w_{l,k} = \mathbf{T}^l_b\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}^w_{b,k-1} \tag{6}$$

Now, we need to obtain the predicted position of the LiDAR points cloud $\boldsymbol{X}^w_{k,j}$ at time k based on $\boldsymbol{P}^w_{l,k}$. In the LiDAR coordinate system, the coordinate of the LiDAR sensor is at the origin. Therefore, the LiDAR points cloud coordinates in the LiDAR coordinate system represent the offset of the points cloud relative to the LiDAR sensor. Therefore, the coordinates of the LiDAR point cloud in the world coordinate system can be obtained by adding $\boldsymbol{P}^w_{l,k}$ and the coordinates of the LiDAR points cloud, as shown in Eq. (7).

$$\mathbf{X}^w_{k,i} = \mathbf{X}^L_{k,i} + \mathbf{P}^w_{l,k} \tag{7}$$

Combining with Eq. (6), the coordinates of the LiDAR points cloud at time k is as shown in Eq. (8).

$$\mathbf{X}^w_{k,i} = \mathbf{X}^L_{k,i} + \mathbf{T}^l_b\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}^w_{b,k-1} \tag{8}$$

By substituting the world coordinate system's point cloud values from Eq. (7) and Eq. (8) into distance formula [13], we can obtain the distance of the LiDAR points cloud to the line (plane) in the world coordinate system. We refer to this as the predicted values and denote it as $d_{\varepsilon p}(d_{\zeta_p})$.

Therefore, the optimization factor of the LiDAR points cloud is shown in Eq. (9).

$$\mathbf{f}_l(\mathbf{z}_l) = \sum_{i=1}^{m}(d_{\varepsilon_m} - d_{\varepsilon_p}) + \sum_{j=1}^{n}(d_{\zeta_m} - d_{\zeta_p}) \tag{9}$$

m and n represent the number of corner points and surface points, respectively.

C. GNSS Factor

For GNSS signal, the pseudo-range can be obtained by multiplying the propagation time by the speed of light v_c. The duration of one pseudocode chip Q_c is known. The propagation time $\boldsymbol{T}_c$ can be obtained by multiplying the number of pseudocode chips N_c by Q_c, so the pseudo-range D_1 is shown in Eq. (10).

$$\boldsymbol{T}_c = N_cQ_c, D_1 = \boldsymbol{T}_cv_c \tag{10}$$

We can also obtain pseudo-range by Eq. (11).

$$D_2 = \|\boldsymbol{P}^i_s - \boldsymbol{P}^i_r\| + O^s_r + T^s_r + I^s_r + M^s_r + v_c\delta_r + \epsilon^s_r \tag{11}$$

where, O^s_r includes orbit errors and clock errors, I^s_r and T^s_r are ionosphere and troposphere errors, M^s_r is the multipath effect, δ_r is the clock error of the receiver and ϵ^s_r is Gaussian white noise. $\boldsymbol{P}^i_s$ and $\boldsymbol{P}^i_r$ denote the Earth-Centered Inertial (ECI) coordinates of the satellite and the receiver, respectively, when the satellite signal arrives at the receiver. The Earth-Centered, Earth-Fixed (ECEF) coordinates of the satellite $\boldsymbol{P}'^e_s$ are included in the ephemeris. The ECEF coordinates of the satellite at the transmission time and at the reception time differ by an angle corresponding to the Earth's rotation. Therefore, we can derive the angle of Earth's rotation, $\theta = \omega_E\boldsymbol{T}_c$. Thus, the ECI coordinate of the satellite at the receiving time can be obtained by Eq. (12):

$$\boldsymbol{P}^e_s = \mathbf{R}_Z(\theta)\boldsymbol{P}'^e_s \tag{12}$$

where $\mathbf{R}_Z(\theta)$ [10] represents a rotation about the z axis of the ECEF frame with magnitude θ.

Hence we have Eq. (13).

$$\|\boldsymbol{P}^i_s - \boldsymbol{P}^i_r\| + O^s_r + T^s_r + I^s_r + M^s_r + v_c\delta_r + \epsilon^s_r = \boldsymbol{T}_cv_c \tag{13}$$

At time $k-1$, the following Eq. (14) holds true.

$$\begin{aligned} D_{1,k-1} =& \boldsymbol{T}_{c,k-1}v_c \\ D_{2,k-1} =& \|\boldsymbol{P}^i_{s,k-1} - \boldsymbol{P}^i_{r,k-1}\| + O^{s,k-1}_r + T^{s,k-1}_r \\ &+ I^{s,k-1}_r + M^{s,k-1}_r + v_c\delta^{k-1}_r + \epsilon^{s,k-1}_r \end{aligned} \tag{14}$$

At time k, there are Eq. (15).

$$\begin{aligned} D_{1,k} =& \boldsymbol{T}_{c,k}v_c \\ D_{2,k} =& \|\boldsymbol{P}^i_{s,k} - \boldsymbol{P}^i_{r,k}\| + O^{s,k}_r + T^{s,k}_r \\ &+ I^{s,k}_r + M^{s,k}_r + v_c\delta^k_r + \epsilon^{s,k}_r \end{aligned} \tag{15}$$

where,

$$\begin{aligned} \boldsymbol{P}^i_{s,k-1} &= \mathbf{R}_Z(\theta_1)\boldsymbol{P}'^e_{s,k-1}, \boldsymbol{P}^i_{r,k-1} = R^i_w\boldsymbol{P}^w_{r,k-1} \\ \boldsymbol{P}^i_{s,k} &= \mathbf{R}_Z(\theta_2)\boldsymbol{P}^{e'}_{s,k}, \boldsymbol{P}^i_{r,k} = R^i_w\boldsymbol{P}^w_{r,k} \\ \theta_1 &= \omega_E\boldsymbol{T}_{c,k-1}, \theta_2 = \omega_E\boldsymbol{T}_{c,k} \end{aligned} \tag{16}$$

In this context, we consider the East-North-Up (ENU) coordinate system as the world coordinate system, and we align the ECI and ECEF coordinate systems at the time of reception. Therefore, $\mathbf{R}^i_w$ is equivalent to $\mathbf{R}^e_n$. According to [10], $\mathbf{R}^n_e$ is shown in Eq. (17).

$$\mathbf{R}^n_e = \begin{bmatrix} -\sin\lambda & -\sin\phi\cos\lambda & \cos\phi\cos\lambda \\ \cos\lambda & -\sin\phi\sin\lambda & \cos\phi\sin\lambda \\ 0 & \cos\phi & \sin\phi \end{bmatrix} \tag{17}$$

λ and ϕ denote the reference point's longitude and latitude.

Then, we obtain the coordinates of the GNSS receiver at moments $k-1$ and k by Eq. (18).

$$\boldsymbol{P}^w_{r,k} = \mathbf{T}^g_b\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}^w_{b,k-1}, \boldsymbol{P}^w_{r,k-1} = \mathbf{T}^g_b\boldsymbol{P}^w_{b,k-1} \tag{18}$$

Then, integrating Eq. (14), (15), (16) and Eq. (18), we can derive Eq. (19).

$$
\begin{aligned}
&\boldsymbol{T}_{c,k-1}v_c - \boldsymbol{T}_{c,k}v_c = \\
&\|\mathbf{R}_Z(\theta_1)\boldsymbol{P}'^{e}_{s,k-1} - \mathbf{R}^i_w\mathbf{T}^g_b\boldsymbol{P}^w_{b,k-1}\| \\
&+ O^{s,k-1}_r + T^{s,k-1}_r + I^{s,k-1}_r \\
&+ M^{s,k-1}_r + v_c\delta^{k-1}_r + \epsilon^{s,k-1}_r \\
&- \|\mathbf{R}_Z(\theta_2)\boldsymbol{P}'^{e}_{s,k-1} - \mathbf{R}^i_w\mathbf{T}^g_b\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}\boldsymbol{P}^w_{b,k-1}\| \\
&- O^{s,k}_r - T^{s,k}_r - I^{s,k}_r \\
&- M^{s,k}_r - v_c\delta^{k}_r - \epsilon^{s,k}_r
\end{aligned}
\tag{19}
$$

We assume that the clock error is equal within a certain period for the same receiver, meaning that $\delta^{k-1}_r = \delta^k_r$. By subtracting the right-hand side from the left-hand side of Eq. (19), we can obtain the optimization factor $\mathbf{f}_g(\mathbf{z}_g)$.

V. The Process of Estimating Spatial Parameters and Position

By combining all the optimization factors from the different sensors, the optimization formulation of the proposed fusion algorithm is as shown in Eq. (20).

$$
\begin{aligned}
\boldsymbol{\xi}^* &= \arg\min_{\boldsymbol{\xi}}\{\|\mathbf{r}_p - \mathbf{J}_p\boldsymbol{\xi}\|^2_{\Sigma_p} + \sum_i \|\mathbf{f}_i(\mathbf{z}_i)\|^2_{\Sigma_i}\} \\
&= \arg\min_{\boldsymbol{\xi}}\{\|\mathbf{r}_p - \mathbf{J}_p\boldsymbol{\xi}\|^2_{\Sigma_p} + \|\mathbf{f}_b(\mathbf{z}_b)\|^2_{\Sigma_b} + \|\mathbf{f}_g(\mathbf{z}_g)\|^2_{\Sigma_g} \\
&\quad + \|\mathbf{f}_v(\mathbf{z}_v)\|^2_{\Sigma_v} + \|\mathbf{f}_l(\mathbf{z}_l)\|^2_{\Sigma_l}\}
\end{aligned}
\tag{20}
$$

where, $\mathbf{z}_i$ represents the measurement data from sensors. $\{\mathbf{r}_p, \mathbf{J}_p\}$ represents the prior information [14].

We calculate the desired variables in two steps, with the state variables represented by Eq. (21) and Eq. (22) respectively.

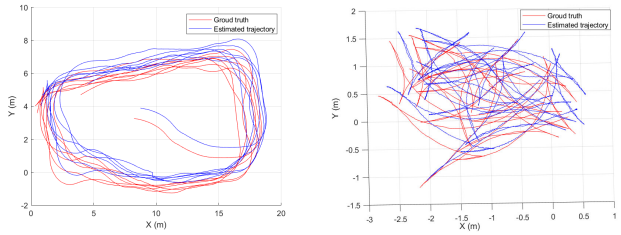
$$
\begin{aligned}
\boldsymbol{\xi} &= \{\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_{n+N_1}\}, \\
\boldsymbol{x}_i &= \{\mathbf{T}^l_b, \mathbf{T}^c_b, \mathbf{T}^g_b, \mathbf{T}_{(\boldsymbol{R},\boldsymbol{t}),i}\}, i \in [n+1, n+N_1]
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
\boldsymbol{\xi} &= \{\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \cdots, \boldsymbol{x}_{n+N_2}\}, \\
\boldsymbol{x}_i &= \{\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t}),i}\} \quad i \in [n+1, n+N_2]
\end{aligned}
\tag{22}
$$

n represent the start of the sliding window. $n+N_1$ and $n+N_2$ represent the end of the sliding window of two stages respectively. N_1 and N_2 represent the number of keyframes in the sliding window.

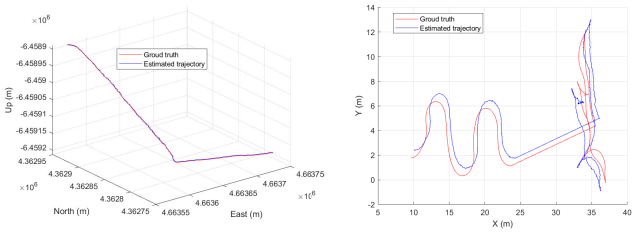
Firstly, we obtain $\mathbf{T}^l_b, \mathbf{T}^c_b, \mathbf{T}^g_b$ through nonlinear optimization algorithm. Second, building upon the results of the first step, we estimate the pose parameters of the vehicle $\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t})}$.

Σ represents the covariance and serves two purposes. Firstly, it normalizes the error factors, allowing them to be combined in a dimensionless manner. Secondly, it represents the expression of error uncertainty, providing a weighted representation of different types of errors. For instance, concerning visual data, we assume that the detected image feature points have an uncertainty of one pixel. By dividing the residual factor of visual data by this covariance, we obtain a dimensionless residual factor. In this paper, we consider the clock errors of the GNSS receiver δ_r at the two moments equal. This way, they cancel out when subtracted in Eq. (19). Regarding the solution methods for nonlinear optimization, there are many existing approaches. We adopt the graph optimization method following the framework in reference [15].



(a) (b)

Fig. 2. (a) Comparison between estimated trajectory using the proposed fusion algorithm and reference trajectory: (a) hall_05; (b) room_dark_06.



(a) (b)

Fig. 3. Comparison between estimated trajectory using the proposed fusion algorithm and reference trajectory: (a) street_06; (b) lift_04.

VI. Simulations And Comparisons

In this section, we utilize publicly available datasets from the Internet to validate the proposed fusion algorithm and conduct comparison with existing algorithms. We utilize the M2DGR dataset [16], a novel large-scale dataset[1] which contains multiple sequences covering a variety of scenarios, including both indoor and outdoor environments. We select sequences street_06, street_07, room_dark_06, lift_04, hall_05.

Fig. 2 and Fig. 3 present the comparative results between the estimated trajectory using the proposed fusion algorithm and the reference trajectory.

From these figures, it can be observed that the proposed algorithm can estimate the position at any time interval. The statistical results of their error are shown in Table. I. The statistical error data about the existing algorithms on these datasets is sourced from [16]. From Table. I, it is evident that due to the individual degradation environments of LiDAR and visual data, some fusion algorithms cannot achieve localization or exhibit significant localization errors that render them impractical. For instance, sequences street_07

[1]https://github.com/SJTU-ViSYS/M2DGR

TABLE I
POSITION ESTIMATION ERROR STATISTICAL RESULTS(RMSE) OF EXISTING FUSION ALGORITHMS AND THE PROPOSED FUSION ALGORITHM ON DIFFERENT DATASETS

Method Sequence	Street_06	Street_07	Hall_05	Lift_04	Room _Dark_06
A-LOAM	0.628	28.940	1.065	1.323	0.314
LeGO LOAM	1.246	35.437	28.940	1.370	0.373
LINS	1.742	12.009	1.010	1.318	2.205
LIO-SAM	0.417	28.642	1.047	1247.153	0.324
ORB3 Pinhole	5.845	none	3.291	none	none
ORB3 Fisheye	95.056	none	none	8.131	none
ORB3 Thermal	30.450	5.927	28.940	2.873	0.404
Cubemap SLAM	98.391	none	5.171	none	none
VINS Mono	124.357	143.725	0.646	5.582	1.001
RTKLIB	6.749	13.096	none	none	none
Proposed	0.1853	2.6354	0.2854	0.5282	0.1486

[1] 'None' indicates initialization failure or inability to determine the location.

and room_dark_06 belong to low illumination environments. Thus, ORB3-Fisheye and ORB3-Pinhole are unable to achieve localization on these datasets. In sequence lift_04, the fusion systems relying solely on visual or LiDAR information, like LIO-SAM with a localization error of 1247.153(m), ORB3-Pinhole, and RTKLIB, cannot achieve localization. Therefore, due to the presence of degraded environments, the localization performance of existing algorithms is relatively poor.

The paper introduces a tightly-coupled odometry framework that integrates visual, LiDAR, inertial, and GNSS data. Its strength lies in harnessing the complementary advantages of different sensors to achieve continuous high-precision localization in complex environments. For instance, in low-light conditions, inertial, LiDAR, and GNSS data can be fully utilized for localization. In the environments where there are significant changes in terrain, the performance of Lidar SLAM will decrease, visual, inertial, and GNSS data can be fused for localization. In GNSS-unfriendly environments, visual, inertial, and LiDAR data can be utilized for localization.

The advantage of this fusion framework lies in fully exploiting multiple sensors to adapt to complex environments. Therefore, compared to existing fusion algorithms based on LiDAR or vision, this algorithm can achieve continuous and high-precision localization in complex environments. Consequently, compared to the existing algorithms, the proposed fusion algorithm is better equipped to adapt effectively to complex environments.

Based on the data in Table. I, the proposed fusion algorithm can improve the positional accuracy by 53.88%, 55.53%, 55.82%, 59.92% and 52.68% on sequences street_06, street_07, hall_05, lift_04 and room_dark_06, respectively, compared to the state of the art algorithms. Although the different methods used for comparison in Table I considered different subset of sensors, it highlights the benefit of the proposed sensor fusion that is capable of utilizing all available sensor data to get a superior performance with higher position accuracy.

VII. CONCLUSIONS

This paper proposes a tightly-coupled algorithm for fusing GNSS, visual, LiDAR, and inertial data based on non-linear optimization for complex environments. Experimental results indicate that the proposed algorithm is better suited for complex environments compared to state-of-the-art algorithms.

REFERENCES

[1] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, "Super odometry: IMU-centric LiDAR-Visual-Inertial estimator for challenging environments," in *Proc. IEEE/RSJ Int. Conf. Intell.Robots Sys*, 2021, pp. 8729–8736.

[2] T. Shan, B. Englot, C. Ratti, and D. Rus, "LVI-SAM: Tightly-coupled LiDAR-Visual-Inertial odometry via smoothing and mapping," in *Proc. IEEE Int. Conf.Robot. Autom*, 2021, pp. 5692–5698.

[3] R. Sun, J. Wang, Q. Cheng, Y. Mao, and W. Y. Ochieng, "A new IMU-aided multiple GNSS fault detection and exclusion algorithm for integrated navigation in urban environments," *GPS Solutions*, vol. 25, pp. 1–17, 2021.

[4] X. Hu, D. Olesen, and P. Knudsen, "Toward high-quality magnetic data survey using UAV: development of a magnetic-isolated vision-based positioning system," *GPS Solutions*, vol. 25, pp. 1–17, 2021.

[5] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV," in *Proc. IEEE Int. Conf.Robot. Autom*, 2014, pp. 4974–4981.

[6] T. Li, H. Zhang, Z. Gao, X. Niu, and N. El-Sheimy, "Tight fusion of a monocular camera, MEMS-IMU, and single-frequency multi-GNSS RTK for precise navigation in GNSS-challenged environments," *Remote Sensing*, vol. 11, no. 6, p. 610, 2019.

[7] W. Wen, T. Pfeifer, X. Bai, and L.-T. Hsu, "It is time for factor graph optimization for GNSS/INS integration: Comparison between FGO and EKF," *arXiv preprint arXiv:2004.10572*, 2020.

[8] Z. Gong, R. Ying, F. Wen, J. Qian, and P. Liu, "Tightly coupled integration of GNSS and vision SLAM using 10-dof optimization on manifold," *IEEE Sens. J.*, vol. 19, no. 24, pp. 12 105–12 117, 2019.

[9] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular Visual-Inertial state estimator," *IEEE Trans. Robot*, vol. 34, no. 4, pp. 1004–1020, 2018.

[10] S. Cao, X. Lu, and S. Shen, "GVINS: Tightly coupled GNSS–Visual–Inertial fusion for smooth and consistent state estimation," *IEEE Trans. Robot*, vol. 38, no. 4, pp. 2004–2021, 2022.

[11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.

[12] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for Visual, Visual-Inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[13] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time." in *RSS.*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[14] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, 2010.

[15] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g 2 o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 3607–3613.

[16] J. Yin, A. Li, T. Li, W. Yu, and D. Zou, "M2DGR: A multi-sensor and multi-scenario SLAM dataset for ground robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2266–2273, 2021.

Authors' Responses to Reviewers' Comments for paper 56716,
"GVIL: Tightly Coupled GNSS-Visual-Inertial-Lidar for Position Estimation in Challenging Environments"

by Yanfang Shi, Baowang Lian, Yonghong Zeng, Ernest Kurniawan and Yugang Ma.

The authors are extremely grateful to the chairs and the reviewers for their careful reading of our manuscript and their helpful comments. We have incorporated these comments and made necessary changes with the highlights in blue typeface in the revision. Please see the detailed responses in this document and the revised manuscript. Hopefully, the version is acceptable for publication. Below are our responses to the reviewers' comments in the exact order they were listed in the reviews.

Responses to chairs:

Q0.0: *The reviewer(s) have suggested some major revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.*

R0.0: The authors appreciate the chairs for giving us a chance to revise the manuscript. According to your suggestions, we have addressed the reviewers' comments and revised our manuscript. Below please fine the responses to reviewers' comments point-to point.

Responses to Reviewer 1:

Q1.0: *Minor grammatical errors and awkward phrasings are present throughout the document. A thorough proofreading is required to enhance the paper's professionalism and readability..*

R1.0: Thanks for thoroughly reviewing the manuscript. We have proofread the entire paper and made accordingly modifications highlighted in blue in the new manuscript.

Responses to Reviewer 2:

Q2.0: *The paper proposes a tight data fusion algorithm that combines inertial, camera, lidar and GNSS information. The results, based on publicly available datasets, are promising when compared to other methods. However, the comparison is often not fair as the methods the proposed algorithm is compared to do not use the same sensors, e.g., it is quite obvious that a GNSS only approach will not work indoors. It would be nice to see a comparison to methods that use the same set of sensors.*

R2.0: Thank you for your careful review of our manuscript. We agree with the reviewer that the other methods used for comparison considered different set of sensors. The contribution of

our sensor fusion approach is that it allows us to use all available sensor data to achieve better positioning accuracy, without causing any degradation compared to the state of the art methods which uses only the subset of sensors. We have highlighted this point in Section VI of the revised manuscript. We would also like to emphasize that in this paper we propose a tightly-coupled odometry framework that integrates visual data, LiDAR data, inertial data, and GNSS data. Its strength lies in leveraging the complementary advantages of different sensors to achieve continuous high-precision localization in complex environments. For instance, in low-light conditions, inertial, LiDAR, and GNSS data can be fully utilized for localization. In the environments where there are significant changes in terrain, the performance of Lidar SLAM will decrease, visual, inertial, and GNSS data can be fused for localization. In GNSS-unfriendly environments, visual, inertial, and LiDAR data can be utilized for localization.

The advantage of this fusion framework lies in fully exploiting multiple sensors to adapt to complex environments. In the experimental section, we compared the position estimation errors of our framework with some other popular odometry frameworks using the same dataset as input. The dataset we used includes GNSS, visual, LiDAR, and inertial data. Some of the compared odometers lack the ability to fuse all the data, so they only utilize certain types of data. Some odometry frameworks failed to localize, and some exhibited significant localization errors. In contrast, our proposed tightly-coupled framework achieved continuous localization with smaller errors, demonstrating its effectiveness in complex environments. The advantage of the algorithm proposed in this paper lies in its greater robustness in complex and challenging environments. Therefore, under the same dataset, the proposed algorithm achieves higher accuracy in position estimation compared to other algorithms. This sufficiently demonstrates that the proposed algorithm is more adaptable to complex environments than existing algorithms, thus reflecting the essence of this paper.

Q2.1: *The presentation needs to be significantly improved. The notation used in the paper is not introduced anywhere and most of the variables are never really explained or introduced, e.g., in (21) what is x? Furthermore, the wording is sometimes also not really clear, e.g., what are spatial parameters?*

R2.1: Thanks for this comment. We have carefully check the entire manuscript and added the introduction statement for each symbol. In Formula 21, $\boldsymbol{x}$ represents the set of variables to be determined, and spatial parameters refer to the relative positions among different sensors. In the first step, $\boldsymbol{x}$ includes $\mathbf{T}_b^l, \mathbf{T}_b^c, \mathbf{T}_b^g, \mathbf{T}_{(\boldsymbol{R},\boldsymbol{t}),i}$. In the second step, $\boldsymbol{x}$ includes $\mathbf{T}_{(\boldsymbol{R},\boldsymbol{t}),i}$. $\mathbf{T}_b^c$, $\mathbf{T}_b^l$, $\mathbf{T}_b^g$ are the spatial parameters between the inertial data and visual (camera) data, LiDAR data, and GNSS data, respectively. Additionally, we have provided detailed explanations for some other variables in the formula, as highlighted in the text.

Responses to Reviewer 3:

Q3.0: *The topic of the paper (positioning in challenging environments) is a perfect match for this conference.*

R3.0: Thank you for your careful review of our manuscript.

Q3.1: *Tight coupling of sensors (GNSS/IMU/LiDAR/Camera) for positioning purposes is well established in the literature. The idea of individual optimization factors is interesting, but many of the tools used are already well-established. The methodology used does not take the different noise levels of each sensor into consideration while fusing them, unlike traditional approaches like Kalman filtering, which weights the measurements based on their noise covariances.*

R3.1: Thank you for the comment. The cost function derived in this paper is based on the assumption that the measurement noise follows Gaussian white noise. Therefore, the measurement noise in this paper follows a Gaussian distribution, and the variances Σ of each sensor's measurements are detailed in the text as shown below, and highlighted in the text.
It is essential to clarify that Σ represents the covariance and serves two purposes. Firstly, it normalizes the error factors from different types of sensor data, allowing them to be combined in a dimensionless manner. Secondly, it represents the expression of error uncertainty, providing a weighted representation of different types of errors. In the proposed fusion system, we establish the uncertainty of errors based on each type of sensor data. For instance, concerning visual data, we assume that the detected image feature points have an uncertainty of about one pixel, which means a variance of one pixel. By dividing the residual factor of visual data by this covariance, we obtain a dimensionless residual factor that can be used in the optimization process. The same method is also applied to other types of sensor data.

Q3.2: *Formalizing the optimization factors for the individual sensors can inspire other researchers to develop their functions. Yet, the paper did not add substantial insights or conclusions that can benefit the field at large.*

R3.2: Thank you for the comment. I have rewritten the innovation points of this paper, as highlighted in the text. The major innovation of this paper lies in establishing optimization factors for GNSS data using GNSS raw data. This has two advantages. Firstly, each piece of GNSS raw data can participate in the optimization process, making the fusion system more adaptable to complex environments. Secondly, the positioning method involving GNSS raw data belongs to absolute positioning, which can eliminate trajectory drift in relative positioning systems to some extent. This opens up a new avenue for research in this field, namely, using GNSS raw data to eliminate trajectory drift without necessarily resorting to complex loop-closure algorithms.

Q3.3: *The paper is very hard to follow for two main reasons: 1) The flow is incohesive and the ordering of ideas is not clear. 2) Many of the mathematical notations are not defined.*

R3.3: Thank you for the comment. We have restructured the layout of the article, rewritten the

innovation points and problem formulation. In the optimization factor chapter, we have also provided a new description, as highlighted in the article. We have provided detailed explanations for the mathematical symbols that were previously undefined, as highlighted in the text.

Q3.4: *Although the paper's topic is a great match for this conference, the novelty and contributions of the work are incremental. Additionally, the paper is tough to follow in its current form. A lot of restructuring should be done to make it easy for the reader to follow the logic of the paper.*

R3.4: Thank you for the comment. As highlighted in the response R3.3 above, we have restructured the layout of the article to improve the flow of presentation. We have also added descriptions to the variables and explanations to the concept of optimization factors and how it facilitate the fusion of multiple sensor data in the optimization problem, leading to a better positioning accuracy. We believe the revised manuscript has a better flow of logic and easier for the reader to follow.

Q3.5: *The paper showcases a tightly coupled sensor fusion scheme for GNSS, visual, LiDAR, and IMU measurements for positioning purposes. The authors derived various optimization factors for each of the sensors. A maximum A Posteriori (MAP) estimation scheme was then proposed to fuse these factors and estimate the position of the user/carrier.* ***Strengths****: The optimization factors approach is interesting;* ***Weaknesses****: Weak introduction and literature review. The literature should be criticized to show the gap this paper is trying to fill; Very incoherent flow and language, hard to follow; Many unidentified mathematical symbols which make the paper very hard to read; Very short results section with very few insights on why the proposed method outperformed other methods.*

R3.5: Thank you for the comment. Fusion systems based on LiDAR or vision are constrained by environmental conditions, causing performance degradation in complex environments and rendering them unable to adapt to challenging environments. Researchers have thus begun considering the integration of GNSS data into fusion computations. However, when GNSS data participates in computations in a loosely-coupled manner, its performance still suffers from unfriendly GNSS signal environments, thus failing to adapt to challenging environments. The robustness of the fusion framework using filtering algorithms to tightly couple GNSS raw data still cannot meet the requirements in challenging environments. The algorithm proposed in this paper, based on nonlinear optimization of tightly coupled GNSS raw data, is precisely developed to address this current situation.

We have also reorganized the layout of the paper and provided detailed explanations for some mathematical symbols that were previously undefined.

Additionally, in the experimental section of the paper, we emphasized why the proposed method outperforms some existing odometry methods, as highlighted in the text. The paper proposes a tightly-coupled odometry framework that integrates visual data, LiDAR data, inertial data, and

GNSS data. Its strength lies in leveraging the complementary advantages of different sensors to achieve continuous high-precision localization in complex environments. For instance, in low-light conditions, inertial, LiDAR, and GNSS data can be fully utilized for localization. In the environments where there are significant changes in terrain, the performance of Lidar SLAM will decrease, visual, inertial, and GNSS data can be fused for localization. In GNSS-unfriendly environments, visual, inertial, and LiDAR data can be utilized for localization. The advantage of this fusion framework lies in fully exploiting multiple sensors to adapt to complex environments. Therefore, compared to existing fusion algorithms based on LiDAR or vision, this algorithm can achieve continuous and high-precision localization in complex environments.

Q3.6: *General comments: Introduction: do not mention the name of the authors.*

R3.6: Thank you for the comment. We have removed all author names from the references as requested, as highlighted in the document.

Q3.7: *It is better to write the full name followed by the acronym not the other way around. Example: unscented Kalman filter (UKF) instead of UKF (Unscented Kalman Filter).Also, note that the name should not have any uppercase letters except if it is usually written this way, e.g., Kalman.*

R3.7: Thank you for the comment. We have checked the entire article and placed all full names in front, with abbreviations in parentheses as shown below, and highlighted them in the text. Oriented FAST and Rotated BRIEF (ORB);Light Detection and Ranging (LiDAR);Global Navigation Satellite System (GNSS);Extended Kalman Filter (EKF);Earth-Centered Inertial (ECI);Earth-Centered, Earth-Fixed (ECEF);East-North-Up (ENU).

Q3.8: *The paper will really benefit from a system model section where you introduce the mathematical notations of the problem and the individual systems (GNSS, camera, LiDAR, and IMU).*

R3.8: Thank you for the comment. We have defined all mathematical symbols in the text that were not previously defined, as shown in the highlighted section.

Q3.9: *The optimization factors of the various sensors are of different nature; hence, their value ranges will differ. For instance, the visual factor comprises distances between pixels while the LiDAR factor comprises differences between distances. So, some sort of normalization of these factors should be taken into consideration. Also, according to my understanding, the proposed method gives all sensors an equal say in the positioning process, regardless of their noise level, which is not optimal. The authors are encouraged to investigate ways to properly weigh the different factors based on how healthy each sensor is compared to others.*

R3.9: Thank you for the comment. The cost function derived in this paper is based on the assumption that the measurement noise follows Gaussian white noise. Therefore, the measurement noise in this paper follows a Gaussian distribution, and the variances Σ of each sensor's measurements

are detailed in the text as shown below, and highlighted in the text.
It is essential to clarify that Σ represents the covariance and serves two purposes. Firstly, it normalizes the error factors from different types of sensor data, allowing them to be combined in a dimensionless manner. Secondly, it represents the expression of error uncertainty, providing a weighted representation of different types of errors. In the proposed fusion system, we establish the uncertainty of errors based on each type of sensor data. For instance, concerning visual data, we assume that the detected image feature points have an uncertainty of about one pixel, which means a variance of 1 pixel. By dividing the residual factor of visual data by this covariance, we obtain a dimensionless residual factor that can be used in the optimization process. The same method is also applied to other types of sensor data.

Q3.10: *Errors: Abstract, Simulation results indicate that using, add a comma after that Index terms: order alphabetically First page, second column, "tightly-couple section: [7], [8]" − >[7] and [8]. Mathematical symbols not defined throughout the paper. Table 1 title: RMSE(m) instead of RMSE(M).*

R3.10: Thank you for the comment. We have made the corresponding modifications in the article as requested, as highlighted in the text.