IMU-Assisted Direct Visual-Laser Odometry in Challenging Outdoor Environments

Quang-Ha Pham[1,2], Ngoc-Huy Tran[1,2*], and Thien-Dao Nguyen[1,2]

[1] Department of Control & Automation, Faculty of Electrical & Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
{pqha.sdh20, tnhuy, dao.nguyen.thien}@hcmut.edu.vn
[2] Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

Abstract. Pose estimation is one of the fundamental capabilities that must be fulfilled by autonomous vehicles prior to performing other tasks such as collision avoidance and motion control. Due to the complexity of outdoor environments, this problem can only be effectively solved by fusing multiple modalities such as LiDAR and camera. In this paper, we propose a novel method to tightly couple inertial measurements from IMU into the emerging direct visual-laser odometry framework. To be more specific, a 2-step optimization-based approach is employed. Firstly, inertial measurements are used to introduce additional constraints in direct image alignment. The estimated pose is then refined in IMU-assisted windowed refinement. To validate the proposed method, we carry out intensive experiments in two recent and challenging datasets: UrbanLoco and USVInland. Experimental results show that our framework enjoys more robust and accurate pose estimation in challenging scenarios compared to that of existing popular methods.

Keywords: Pose Estimation, Visual Odometry, SLAM, Sensor Fusion

1 Introduction

Pose estimation is crucial for robotic navigation. In robotics research, this task can also be named visual odometry, laser odometry, or more generally, simultaneous localization and mapping (SLAM). Due to its widespread applications ranging from autonomous driving [1] to virtual reality [2], an enormous effort has been made in both academia and industry to realize a highly accurate, robust, and efficient implementation of pose estimation for long-term localization in large-scale and dynamic environments [3].

In the literature, there exists a large variety of methods for pose estimation, most of which rely on RGB-D cameras or LiDARs. Current state-of-the-art indirect methods, either visual-based [4][5] or laser-based [6][7], focus on enhancing the robustness of feature extraction while employing a common backend optimization framework. Recently, there has been steady attention towards featureless methods which directly utilize raw sensory data to estimate motion, such as the visual-based DSO [8] or laser-based SUMA [9]. The main advantage of these classes of direct methods lies in their coherent integration of the delicate data association process into the nonlinear optimization module, thus being capable of pushing the limit of pose estimation in some very challenging outdoor environments. However, direct formulation of pose estimation are much more non-convex, thus requiring more careful tweaks in every optimization step to ensure proper convergence.

To ease the development of direct pose estimation, some authors suggest fusing different modalities either visual-inertial-based [10], laser-inertial-based [11], or visual-laser-based [12]. By incorporating multiple sensors, direct methods become more easily implementable thanks to more complementary constraints available and reduction of variables that need to be estimated. In this paper, we follow this strategy and push forwards the development of direct methods by tightly coupling IMU measurements into the direct visual-laser odometry framework DVL-SLAM [12]. More specifically, we propose a novel optimization-based framework that fuses inertial data in 2 steps. Firstly, inertial measurements are used to introduce additional constraints in direct image alignment. The estimated pose is then refined in IMU-assisted windowed refinement. The introduction of IMU is very beneficial since the ill-effects of sparse LiDAR depth association and high sensitivity of image warping are minimized. To highlight the advantages of our framework, we carry out intensive experiments in the challenging UrbanLoco [13] and USVInland [14] datasets since the scenarios when LiDAR or camera produces extremely degraded measurements can be easily encountered in urban streets and riverine waterways.

In summary, our contributions are:

- A direct and tightly-coupled visual-laser-inertial odometry system.
- Additional constraints from IMU measurements to aid direct image alignment.
- A model to incorporate inertial constraints into windowed refinement.
- Thorough evaluation in UrbanLoco and USVInland datasets to show that our method outperforms current state-of-the-art ones.

2 Framework

2.1 Notation

Let ${}^{i}\mathbf{p}_k \in \mathbb{R}^2$ and ${}^{i}d_k \in \mathbb{R}$ be image coordinate and inverse depth of a candidate point k in frame i respectively. Let $\mathbf{T}_i, \mathbf{T}_j \in \mathrm{SE}(3)$ be IMU's poses when frame i and j are captured respectively. Each pose is composed of a rotation matrix and a translation vector:

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ \mathbf{0} & 1 \end{bmatrix}, \qquad \mathbf{T}_j = \begin{bmatrix} \mathbf{R}_j & \mathbf{t}_j \\ \mathbf{0} & 1 \end{bmatrix} \tag{2.1}$$

Let $\mathbf{T}_{\mathrm{CI}} \in \mathrm{SE}(3)$ be the extrinsic transformation from the IMU's to camera's frame of reference. Then the corresponding coordinate ${}^{j}\mathbf{p}_k \in \mathbb{R}^2$ of the candidate point k in frame j is computed by the following formula:

$$ {}^{j}\mathbf{p}_k = \kappa(\mathbf{T}_{\mathrm{CI}}\mathbf{T}_j^{-1}\mathbf{T}_i\mathbf{T}_{\mathrm{CI}}^{-1} \cdot \kappa^{-1}({}^{i}\mathbf{p}_k, {}^{i}d_k)) \tag{2.2} $$

where $\kappa(\cdot)$ and $\kappa^{-1}(\cdot)$ are the camera pinhole projection and back-projection models respectively. $< \cdot >$ is the action operator in SE(3) [15]. Define $I_i, I_j : \mathbb{R}^2 \mapsto \mathbb{R}$ be the mapping from the pixel's coordinate to its intensity value in frame i and j respectively. As in [8], the photometric residual $r_p \in \mathbb{R}$ between each pair $\{{}^{i}\mathbf{p}_k, {}^{j}\mathbf{p}_k\}$ is defined as:

$$ r_p = (I_j({}^{j}\mathbf{p}_k) - b_j) - \frac{e^{a_j}}{e^{a_i}}(I_i({}^{i}\mathbf{p}_k) - b_i) \tag{2.3} $$

where $a_i, b_i, a_j, b_j \in \mathbb{R}$ are affine brightness parameters that accounts for the deviation from photo-consistent assumption in real-world environments.

Let the preintegrated IMU rotation, velocity and position measurements [16] from frame i at timestamp t_i to frame j at timestamp t_j be $\Delta\mathbf{R} \in \mathrm{SO}(3)$, $\Delta\mathbf{v} \in \mathbb{R}^3$ and $\Delta\mathbf{t} \in \mathbb{R}^3$ respectively. Let ${}^{a}\mathbf{b}_i, {}^{g}\mathbf{b}_i, {}^{a}\mathbf{b}_j, {}^{g}\mathbf{b}_j \in \mathbb{R}^3$ be the current accelerometer's and gyroscope's bias estimates when frame i and j are captured respectively. Let ${}^{g}\mathbf{J}_{\Delta R}, {}^{a}\mathbf{J}_{\Delta v}, {}^{g}\mathbf{J}_{\Delta v}, {}^{a}\mathbf{J}_{\Delta t}, {}^{g}\mathbf{J}_{\Delta t} \in \mathbb{R}^{3\times 3}$ be the jacobian matrices with respect to IMU's bias changes between frame i to frame j. Then the inertial residual $\mathbf{r}_s \in \mathbb{R}^{15}$ between the two frames is defined as:

$$ \mathbf{r}_s = \begin{bmatrix}\mathbf{r}_R^{\intercal}\ \mathbf{r}_v^{\intercal}\ \mathbf{r}_t^{\intercal}\ \mathbf{r}_{ba}^{\intercal}\ \mathbf{r}_{bg}^{\intercal}\end{bmatrix}^{\intercal} \tag{2.4} $$

where the component residuals $\mathbf{r}_R, \mathbf{r}_v, \mathbf{r}_t, \mathbf{r}_{ba}, \mathbf{r}_{bg} \in \mathbb{R}^3$ are defined as:

$$ \begin{aligned} \mathbf{r}_R &= \mathbf{R}_j \ominus (\mathbf{R}_i(\Delta\mathbf{R} \oplus {}^{g}\mathbf{J}_{\Delta R}\mathbf{r}_{bg})) \\ \mathbf{r}_v &= \mathbf{R}_i^{\intercal}(\mathbf{v}_j - \mathbf{v}_i - \mathbf{g}\Delta t) - (\Delta\mathbf{v} + {}^{a}\mathbf{J}_{\Delta v}\mathbf{r}_{ba} + {}^{g}\mathbf{J}_{\Delta v}\mathbf{r}_{bg}) \\ \mathbf{r}_t &= \mathbf{R}_i^{\intercal}(\mathbf{t}_j - \mathbf{t}_i - \mathbf{v}_i\Delta t - \frac{1}{2}\mathbf{g}\Delta t^2) - (\Delta\mathbf{t} + {}^{a}\mathbf{J}_{\Delta t}\mathbf{r}_{ba} + {}^{g}\mathbf{J}_{\Delta t}\mathbf{r}_{bg}) \\ \mathbf{r}_{ba} &= {}^{a}\mathbf{b}_j - {}^{a}\mathbf{b}_i \\ \mathbf{r}_{bg} &= {}^{g}\mathbf{b}_j - {}^{g}\mathbf{b}_i \end{aligned} \tag{2.5} $$

Let $\bar{\mathbf{T}}_j, \bar{a}_j, \bar{b}_j, \bar{\mathbf{v}}_j, {}^{a}\bar{\mathbf{b}}_j, {}^{g}\bar{\mathbf{b}}_j$ be prior estimates of $\mathbf{T}_j, a_j, b_j, \mathbf{v}_j, {}^{a}\mathbf{b}_j, {}^{g}\mathbf{b}_j$ respectively. Then the prior residual $\mathbf{r}_o \in \mathbb{R}^{17}$ is defined as:

$$ \mathbf{r}_o = \begin{bmatrix} \mathbf{T}_j \ominus \bar{\mathbf{T}}_j \\ a_j - \bar{a}_j \\ b_j - \bar{b}_j \\ \mathbf{v}_j - \bar{\mathbf{v}}_j \\ {}^{a}\mathbf{b}_j - {}^{a}\bar{\mathbf{b}}_j \\ {}^{g}\mathbf{b}_j - {}^{g}\bar{\mathbf{b}}_j \end{bmatrix} \tag{2.6} $$

2.2 Candidate Point Selection

Upon receiving a new frame, we project the corresponding LiDAR pointcloud into the current image, then follow a selection strategy similar to [12] to get a set of informative pixels with depth which we call candidate points.

2.3 Frame Tracking

When a new frame j is received, its pose is tracked with respect to the current reference frame i by direct image alignment. This can be formulated as an IRLS problem in which the energy to be minimized is as followed:

$$E = E_p + \lambda_s E_s + \lambda_o E_o = \sum w_p r_p^2 + \mathbf{r}_s^\intercal \mathbf{\Sigma}_s^{-1} \mathbf{r}_s + \mathbf{r}_o^\intercal \mathbf{\Sigma}_o^{-1} \mathbf{r}_o \tag{2.7}$$

and the optimizing variale, expressed in a composite manifold [15], as followed:

$$\mathbf{x} = \langle \mathbf{T}_j, a_j, b_j, \mathbf{v}_j, {}^a\mathbf{b}_j, {}^g\mathbf{b}_j \rangle \in \langle \mathrm{SE}(3), \mathbb{R}, \mathbb{R}, \mathbb{R}^3, \mathbb{R}^3, \mathbb{R}^3 \rangle \tag{2.8}$$

In this case, we have introduced additional energy terms from inertial measurements to better constrain gradient steps in the traditional direct image alignment. In particular, while inertial term E_s ensures that our system catches up with abrupt movements, prior term E_o provides further guarantee that the estimated result is not too optimistic. These two terms are relatively weighted with respect to E_p by means of weighting factors λ_s and λ_o.

For computational efficiency, we utilize the inverse compositional scheme and alter the photometric residual as followed:

$$r_p = (I_i({}^i\mathbf{p}_k) - b_i) - \frac{e^{a_i}}{e^{a_j}}(I_j({}^j\mathbf{p}_k) - b_j) \tag{2.9}$$

This prevents the system from recomputing the photometric jacobians at each iteration of the gradient steps, thus improving the overall speed.

To minimize the effect of outliers, the weight w_p of each residual is derived from the residual t-distribution whose expectation μ, standard deviation σ and degree of freedom υ are chosen as followed:

$$\begin{aligned} \mu &= \mathrm{median}\{r_p\} \\ \sigma &= 1.4826\,\mathrm{mad}\{r_p - \mu\} \\ \upsilon &= \frac{4\,\mathrm{kurtosis}\{r_p\} - 6}{\mathrm{kurtosis}\{r_p\} - 3} \end{aligned} \tag{2.10}$$

For an optimal tuning, similar to [17], we further minimize the negative log likelihood of the probability density function using Nelder-Mead method to get the best fitted residual distribution. Then the residual weight can be calculated as followed:

$$w_p = \frac{\upsilon + 1}{\upsilon + (\frac{r_p - \mu}{\sigma})^2} \tag{2.11}$$

For inertial term, $\mathbf{\Sigma}_s \in \mathbb{R}^{15\times15}$ is obtained from preintegrating the IMU covariances from frame i to frame j. In case of prior term, the estimated value and inverse of hessian of the last tracking phase are used for the prior estimate and prior covariance $\mathbf{\Sigma}_o \in \mathbb{R}^{17\times17}$ of the current one.

Constant velocity model is used to obtain the initial guess at the beginning of the tracking. To account for large displacement between frames, we follow the coarse-to-fine pyramid scheme [8]: create a set of downsized images, then the estimated value at the current level acts as an initial guess for the subsequent one.

2.4 Keyframe Management

Let N_p be number of candidate points in the reference frame. For new reference frame creation, similar to [8], we combine four criteria as followed:

- New reference frame is needed if the field of view changes, which is quantified by the optical flow from the reference frame to the current frame:

$$f_1 = \sqrt{\frac{1}{N_p}\sum_{k=1}^{N_p} \|{}^i\mathbf{p}_k - {}^j\mathbf{p}_k\|^2} \tag{2.12}$$

- Occlusion occurs more frequently when camera experiences significant translational movement, which is measured by the optical flow without rotation:

$$f_2 = \sqrt{\frac{1}{N_p}\sum_{k=1}^{N_p} \|{}^i\mathbf{p}_k - {}^j\mathbf{p}_k{}'\|^2} \tag{2.13}$$

  where the rotational part of $\mathbf{T}_j^{-1}\mathbf{T}_i$ is set to identity.
- Large camera exposure that heavily violates the photo-consistent assumption also requires a new reference frame:

$$f_3 = |a_j - a_i| \tag{2.14}$$

- Large camera movement increases the uncertainty of IMU preintegration. Therefore, we introduce a new criterium:

$$f_4 = \operatorname{tr} \mathbf{\Sigma}_s \tag{2.15}$$

Finally, if $w_1 f_1 + w_2 f_2 + w_3 f_3 + w_4 f_4 > 1$ then the current frame becomes a new reference one. The four parameters w_1, w_2, w_3, w_4 represent the relative contribution of each indicator and are manually tuned based on circumstances.

2.5 Windowed Refinement

After creating a new reference frame, we perform windowed refinement by concurrently optimizing states of a fixed number of some most recent reference frames called keyframes. Mathematically speaking, this leads to solving an IRLS problem with the energy:

$$E = E_p + \lambda_s E_s = \sum w_p r_p^2 + \lambda_s \sum \mathbf{r}_s^\intercal \mathbf{\Sigma}_s^{-1} \mathbf{r}_s \tag{2.16}$$

and the optimizing variable, expressed in a composite manifold [15], as followed:

$$\mathbf{x} = \langle \mathbf{T}_{1:N_f}, a_{1:N_f}, b_{1:N_f}, \mathbf{v}_{1:N_f}, {}^a\mathbf{b}_{1:N_f}, {}^g\mathbf{b}_{1:N_f} \rangle \tag{2.17}$$

As can been seen in Figure 1, we have introduced a model that not only preserves dense photometric connections but also leaves space for inertial connections between keyframes. While photometric energy terms are created when

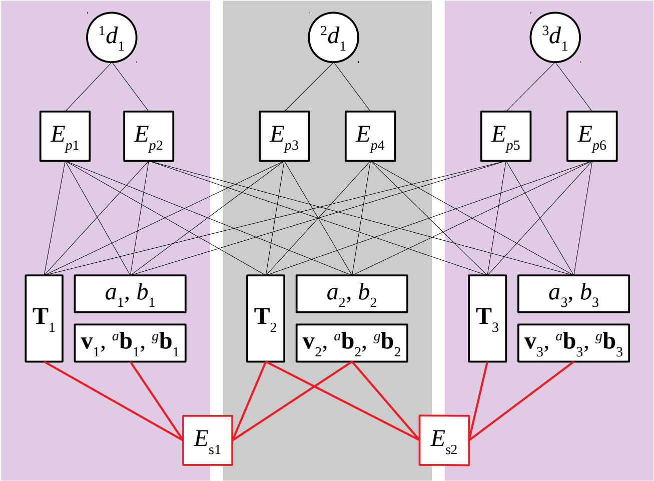


Fig. 1: Energy terms in windowed refinement.

a set of candidate points from one keyframe is projected into any other keyframe in the current sliding window, inertial energy terms are generated by preintegrating all measurements between adjacent keyframes. All keyframe's states are then jointly solved in a nonlinear iterative procedure similar to that of frame tracking for their optimal values. Note that because of the large state space of this problem, we implement an approximate linearization at every gradient step to speed up the computational speed but still guarantee correct convergence.

3 Results

The proposed framework is validated in two challenging public datasets in which LiDAR, camera, and IMU can be used concurrently. For comparison, we reimplement DVL-SLAM [12] with loop closure turned off by disabling IMU integration in our framework. All experiments are run by a PC with Intel Core i5-11400F and 32GB RAM.

3.1 Metrics Selection

To quantify the framework's performance, two different metrics systems are used. Let $\{\mathbf{P} \in \mathrm{SE}(3)\}$ and $\{\mathbf{Q} \in \mathrm{SE}(3)\}$ be the estimated trajectory and ground truth respectively. To evaluate the global consistency of the estimated trajectory, ATE (Absolute Trajectory Error) [18] for each synchronized pose pair are calculated as followed:

$$\mathrm{ATE}_i = \mathbf{Q}_i^{-1}\mathbf{S}\mathbf{P}_i \tag{3.1}$$

where $\mathbf{S} \in \mathrm{SE}(3)$ is the rigid-body transformation that aligns the estimated trajectory to ground truth (set to identity in the remaining article). Then rootmean-squares ATE (RMSE ATE) are used for overall trajectory evaluation:

$$\mathrm{RMSE}(\mathrm{ATE}_{1:n}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathrm{trans}(\mathrm{ATE}_i)\|^2} \quad [\mathrm{m}] \tag{3.2}$$

where trans(·) extracts the translational vector of the rigid-body transformation.

To evaluate the local consistency of the estimated trajectory, similar to [18], RPE (Relative Pose Error) for each pose pair is used:

$$\mathrm{RPE}_i = (\mathbf{Q}_i^{-1}\mathbf{Q}_{i+\Delta_i})^{-1}(\mathbf{P}_i^{-1}\mathbf{P}_{i+\Delta_i}) \tag{3.3}$$

where Δ_i is selected so that $\mathbf{Q}_i$ and $\mathbf{Q}_{i+\Delta_i}$ is separated by a fixed δ distance. However, for overall trajectory evaluation, we follow [19] and separate translational and rotational parts of RPEs over all pose pairs:

$$\begin{aligned} \mathrm{trans}(\mathrm{RPE}_{1:n}, \delta) &= \frac{1}{s}\frac{1}{n}\sum_{i=1}^{n} \|\mathrm{trans}(\mathrm{RPE}_i)\|^2 \times 100 \quad [\%] \\ \mathrm{rot}(\mathrm{RPE}_{1:n}, \delta) &= \frac{1}{s}\frac{1}{n}\sum_{i=1}^{n} \mathrm{rot}(\mathrm{RPE}_i) \quad [\mathrm{deg/m}] \end{aligned} \tag{3.4}$$

where rot(·) extracts the angle of rotation (angle-axis representation) of the rigid-body transformation and s is the overall travelled distance:

$$s = \sum_{i=1}^{n-1} \left\|\mathrm{trans}(\mathbf{Q}_i^{-1}\mathbf{Q}_{i+1})\right\|^2 \tag{3.5}$$

3.2 Evaluation on USVInland

This dataset represents an unmanned surface vessel (USV) travelling along a waterway area, which exhibits challenges to odometry system due to existence of vegetated scenes and absorption of laser points from the water surface as in Figure 2. In this dataset, we utilize the left camera and the IMU inside the camera

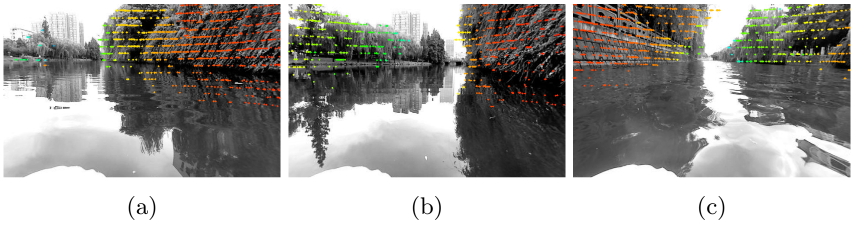
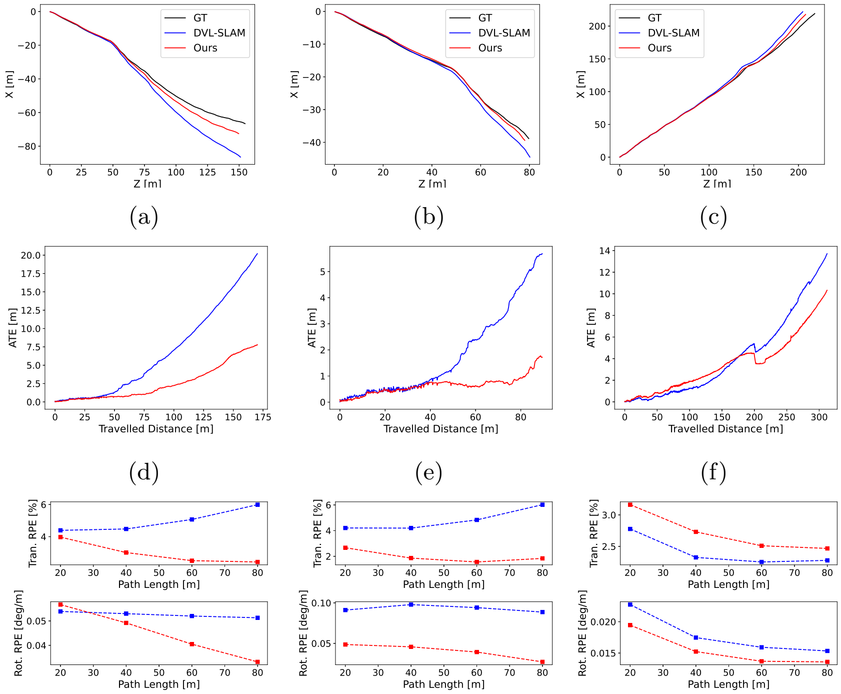

(a) (b) (c)

Fig. 2: Selected scenes from USVInland dataset. Images (a), (b) and (c) with candidate points overlaid (hotter means nearer) are extracted from sequence N03_3_605_760, N03_4_440_523 and N03_5_12_340 respectively.

for estimation. All raw visual, laser, and inertial measurements are synchronized before being fed into the system. For ground truth generation, we first transform all GNSS data and external IMU measurements into a common reference frame, then synchronize with LiDAR's pointcloud and interpolate for missing data.

Our results in this dataset are shown in Figure 3. It can be easily seen that our estimated path aligns more closely to the ground truth than that of DVL-SLAM, which leads to our ATEs and RPEs being significantly lower. This implies that our framework exhibits more global accurancy due to higher ATEs and local accurancy due to higher RPEs. As IMU introduces strict rotational constraints, our rotational RPEs are consistently lower in all sequences, thus reducing the overall rotational drift. In sequence N03_5_12_340, the fact that our framework exhibits slightly poorer translational RPEs may stem from lengthy travelled path and jerky movements at the start of the journey.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3: Results from USVInland dataset. Each column shows plots from sequences N03_3_605_760, N03_4_440_523, and N03_5_12_340 respectively. Each row from top to bottom shows the estimated paths, ATEs and RPEs respectively.

3.3 Evaluation on UrbanLoco

In this dataset, a car is manually driven along urban streets, which poses a great challenge to odometry system due to road elevation and dynamic objects as in Figure 4. The frontal camera and consumer-grade Xsens IMU are used in our framework. Similar to USVInland, all measurements from LiDAR, camera, and IMU are synchronized before being processed by the system. The estimated trajectory is then aligned with 6-DOF ground truth generated from a precise GNSS-INS system for evaluation purposes.

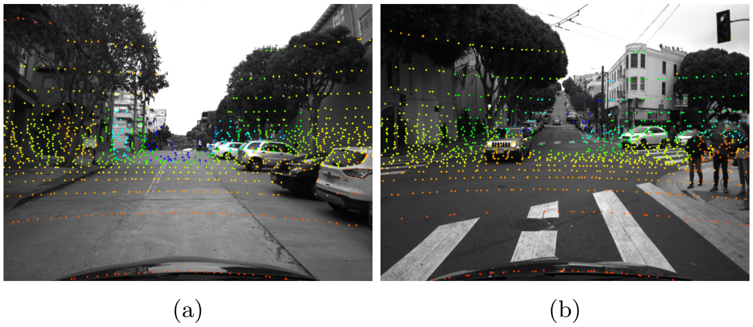
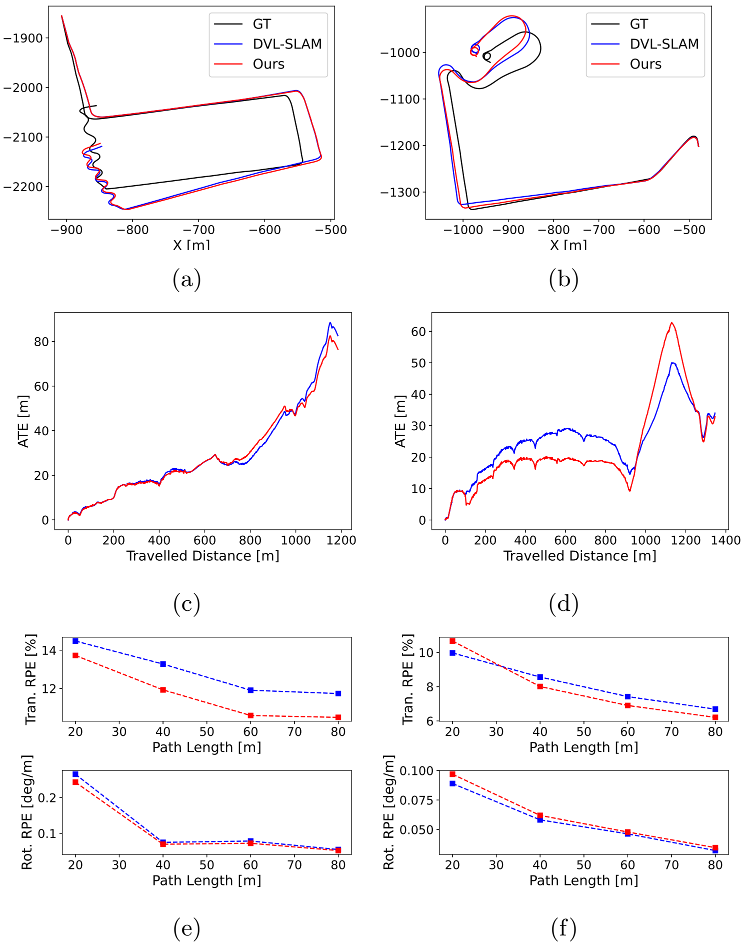


Fig. 4: Selected scenes from UrbanLoco dataset. Image (a) and (b) with candidate points overlaid (hotter means nearer) is extracted from sequence CALombardStreet20190828190411 and CAColiTower20190828184706 respectively.

Our results in this dataset are shown in Figure 5. It is easily noticed that our framework only performs slightly better than DVL-SLAM in all sequences. This stems from the fact that urban streets poccess many challenges that our system cannot fully deal with such as high road elevation and sudden appearance of dynamic objects (cars, pedestrians, etc). In addition, we witness that there exists abrupt illumination changes in some scenes, which significantly contributes to the degradation of visual-based odometry system.



(a) (b)

(c) (d)

(e) (f)

Fig. 5: Results from UrbanLoco dataset. Each column shows plots from sequence CALombardStreet20190828190411 and CAColiTower20190828184706 respectively. Each row from top to bottom shows the estimated paths, ATEs and RPEs respectively.

To summarize, we compare the RMSE ATE of our framework and DVL-SLAM in all sequences in Table 1. Again, ours performs consistently better than DVL-SLAM in all sequences.

Table 1: Comparison of RMSE ATE [m] in all sequences.

	DVL-SLAM	Ours
N03_3_605_760	9.035	**3.388**
N03_4_440_523	2.437	**0.719**
N03_5_12_340	5.155	**4.023**
CALombardStreet20190828190411	44.150	**42.720**
CAColiTower20190828184706	24.854	**24.721**

4 Conclusion

We have presented a multi-sensor odometry system in which LiDAR, camera and IMU are fused in a tight and direct manner. Specifically speaking, an IMU-assisted direct image alignment is introduced to boost the frame tracking performance. The estimated pose is then further refined in a novel model that tightly incorporates inertial constraints between keyframes. Extensive quantitative results demonstrate that our method exhibits better global and local accuracy than the current state of the art that only utilizes laser and visual data.

Acknowledgement

We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

References

1. Bresson, G., Alsayed, Z., Yu, L., Glaser, S.: Simultaneous localization and mapping: A survey of current trends in autonomous driving. IEEE Transactions on Intelligent Vehicles 2(3), 194–220 (2017)
2. Jinyu, L., Bangbang, Y., Danpeng, C., Nan, W., Guofeng, Z., Hujun, B.: Survey and evaluation of monocular visual-inertial slam algorithms for augmented reality. Virtual Reality & Intelligent Hardware 1(4), 386–410 (2019)
3. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics 32(6), 1309–1332 (2016)
4. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics 37(6), 1874–1890 (2021)

5. Cvišić, I., Ćesić, J., Marković, I., Petrović, I.: Soft-slam: Computationally efficient stereo visual simultaneous localization and mapping for autonomous unmanned aerial vehicles. Journal of field robotics 35(4), 578–595 (2018)
6. Zhang, J., Singh, S.: Low-drift and real-time lidar odometry and mapping. Autonomous Robots 41(2), 401–416 (2017)
7. Qin, C., Ye, H., Pranata, C.E., Han, J., Zhang, S., Liu, M.: Lins: A lidar-inertial state estimator for robust and efficient navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 8899–8906. IEEE (2020)
8. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence 40(3), 611–625 (2017)
9. Behley, J., Stachniss, C.: Efficient surfel-based slam using 3d laser range data in urban environments. In: Robotics: Science and Systems. vol. 2018, p. 59 (2018)
10. Von Stumberg, L., Usenko, V., Cremers, D.: Direct sparse visual-inertial odometry using dynamic marginalization. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 2510–2517. IEEE (2018)
11. Xu, W., Cai, Y., He, D., Lin, J., Zhang, F.: Fast-lio2: Fast direct lidar-inertial odometry. arXiv preprint arXiv:2107.06829 (2021)
12. Shin, Y.S., Park, Y.S., Kim, A.: Dvl-slam: sparse depth enhanced direct visual-lidar slam. Autonomous Robots 44(2), 115–130 (2020)
13. Wen, W., Zhou, Y., Zhang, G., Fahandezh-Saadi, S., Bai, X., Zhan, W., Tomizuka, M., Hsu, L.T.: Urbanloco: A full sensor suite dataset for mapping and localization in urban scenes. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 2310–2316. IEEE (2020)
14. Cheng, Y., Jiang, M., Zhu, J., Liu, Y.: Are we ready for unmanned surface vehicles in inland waterways? the usvinland multisensor dataset and benchmark. IEEE Robotics and Automation Letters 6(2), 3964–3970 (2021)
15. Sola, J., Deray, J., Atchuthan, D.: A micro lie theory for state estimation in robotics. arXiv preprint arXiv:1812.01537 (2018)
16. Forster, C., Carlone, L., Dellaert, F., Scaramuzza, D.: On-manifold preintegration for real-time visual–inertial odometry. IEEE Transactions on Robotics 33(1), 1–21 (2016)
17. Zubizarreta, J., Aguinaga, I., Montiel, J.M.M.: Direct sparse mapping. IEEE Transactions on Robotics 36(4), 1363–1370 (2020)
18. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)