

Night-Rider: Nocturnal Vision-aided Localization in Streetlight Maps Using Invariant Extended Kalman Filtering

Tianxiao Gao¹, Mingle Zhao¹, Chengzhong Xu¹, and Hui Kong^{1*}

Abstract—Vision-aided localization for low-cost mobile robots in diverse environments has attracted widespread attention recently. Although many current systems are applicable in daytime environments, nocturnal visual localization is still an open problem owing to the lack of stable visual information. An insight from most nocturnal scenes is that the static and bright streetlights are reliable visual information for localization. Hence we propose a nocturnal vision-aided localization system in streetlight maps with a novel data association and matching scheme using object detection methods. We leverage the Invariant Extended Kalman Filter (InEKF) to fuse IMU, odometer, and camera measurements for consistent state estimation at night. Furthermore, a tracking recovery module is also designed for tracking failures. Experimental results indicate that our proposed system achieves accurate and robust localization with less than 0.2% relative error of trajectory length in four nocturnal environments.

I. INTRODUCTION

With the increasingly versatile applications of mobile robots and autonomous vehicles, accurate and robust localization has been in great demand. Since cameras are cost-effective and informative, extensive visual SLAM methods [1]–[4] have been developed for diverse mobile robots. Nevertheless, dynamic objects make the visual localization without prior maps prone to failures and the online-built local map increases computation costs. Therefore, many visual localization methods [5]–[12] using pre-built maps are proposed recently. These methods can be categorized to feature-based and semantics-based. Feature-based methods [5], [9], [10], [12] extract the low-level features and match them with map points for pose estimation. Semantics-based visual localization methods [6]–[8], [11] utilize the segmented poles, lanes, signboards, and curbs to estimate the 2D-3D correspondences between images and the map. Although they achieve accurate localization, the vast majority of them only consider localization in the daytime or bright scenarios. Currently, the visual localization in nighttime scenes is still an open problem.

One problem of visual localization in nocturnal environments is the drastically changed luminance. The background is illuminated by bright light sources and becomes completely dark in the absence of light. Such a problem makes the current visual localization methods not applicable to nocturnal scenes. For feature-based visual localization



Fig. 1: The proposed nocturnal localization system in the nighttime campus streetlight map. Top-left image shows the robot configuration. Top-right image is the trajectory (red) overlaid with the daytime satellite map for visualization. The middle images display the matches of 2D streetlight detections and 3D streetlight clusters. The bottom image depicts the estimated trajectory in the streetlight map.

methods, low-level feature extraction is not effective at night as it is sensitive to light conditions. The unstable luminance at night makes the semantics-based methods become invalid as well [13]. Fortunately, there are abundant static artificial streetlights that can provide a source of rare perceptive information in most nocturnal scenarios (e.g. nighttime campuses or parks). Additionally, the light saturation makes streetlights easy to detect from images. Therefore, the purpose of this paper is to localize mobile robots in nocturnal environments by leveraging streetlights to establish connections between the camera measurements and the 3D environment.

In this paper, we propose an accurate and robust nocturnal vision-aided localization system in streetlight maps based on an Inertial Measurement Unit (IMU), an odometer, and a monocular camera. Note that NightNav [14] has a similar idea of using streetlights for nighttime visual localization. However, NightNav introduces constraints for pose optimization by assuming that the area of streetlights is proportional to the inverse of the square of the distance between the camera and streetlights, which is not always true as the angle of view, occlusion, and overlap of lights can significantly interrupt the relation. In contrast, in our system, there is no assumption about nocturnal scenes. Besides, the pre-stored images of all streetlights in NightNav drastically increase the storage cost while our system only requires lightweight maps with sparse streetlight points. Fig. 1 shows the robot

*Corresponding author.

¹Tianxiao Gao, Mingle Zhao, Chengzhong Xu, and Hui Kong are with the State Key Laboratory of Internet of Things for Smart City (SKL-IOTSC), Faculty of Science and Technology, University of Macau, Macao, China. ({ga0.tianxiao, zhao.mingle}@connect.umac.mo, {czxu, huikong}@um.edu.mo)

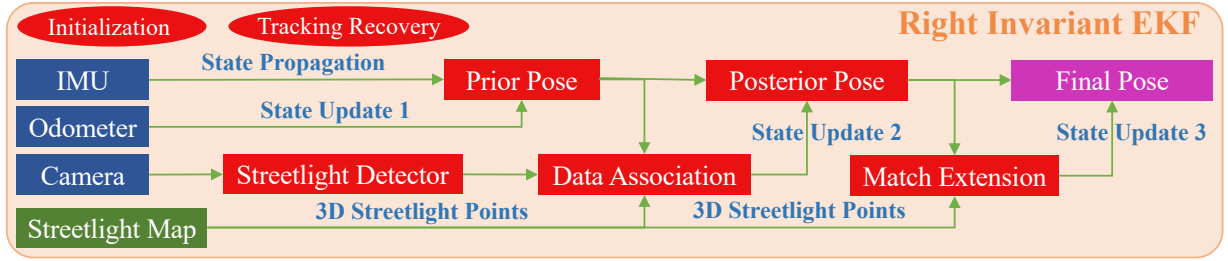


Fig. 2: System overview of the nocturnal vision-aided localization system. The InEKF is utilized for state estimation.

configuration and an example of the estimated trajectory in the streetlight map. The main contributions of the paper are:

- We propose a real-time nocturnal vision-aided localization system, Night-Rider. The system can achieve accurate and robust localization by leveraging limited perceptive information from streetlights at night.
- We propose a novel matching scheme for data association in streetlight maps which enables 6-DoF localization with extremely low drift in nighttime scenarios.
- To mitigate potential errors in nocturnal scenes, the Invariant Extended Kalman Filter (InEKF) is introduced for consistent and stable state estimation.
- We perform extensive experiments on a wheeled robot in real nighttime scenes to validate the effectiveness, accuracy, and robustness of the proposed system.

II. SYSTEM OVERVIEW

Fig. 2 shows the framework of the proposed nocturnal localization system. The streetlight map is constructed using only the 3D streetlight points. We introduce the right InEKF due to its accuracy and consistency in the system for fusing monocular camera, IMU, and odometer measurements. Specifically, the robot state is propagated using IMU measurements, and the odometer measurements are utilized for the first state update. Next, the updated states are fed into the proposed data association module to establish correspondences between 3D streetlight clusters and 2D detections. The matches will be used for the second state update. More 2D-3D streetlight matches are then obtained through the binary segmentation in the match extension module and are applied to the third state update. Additionally, a tracking recovery module is proposed to resolve localization failures resulting from insufficient matches.

As for the notations and frames, $(\cdot)^w$, $(\cdot)^c$, $(\cdot)^o$ and $(\cdot)^b$ denote the world, camera, odometer and body frames. The IMU frame coincides with the body frame. The first IMU frame is regarded as the world frame and the gravity vector \mathbf{g}^w is constant in the world frame. $\mathbf{R}_b^w \in SO(3)$ and $\mathbf{p}_b^w \in \mathbb{R}^3$ are rotation and translation from the world frame to body frame. All velocities in this paper are with respect to the world frame. For example, ${}^w\mathbf{v}_b \in \mathbb{R}^3$ is the velocity of body frame with respect to and represented in the world frame. (\cdot) and $(\hat{\cdot})$ represent the measurement and the estimate, respectively.

III. STATE ESTIMATION AND LOCALIZATION

In this section, we first briefly introduce the generation of streetlight map. Then we illustrate the state representation of

the system followed by the InEKF for state estimation fusing IMU, odometer, and camera measurements. The proposed data association and match extension methods enable the system to accurately localize 6-DoF poses at night. Finally, we explicate the tracking recovery module for tracking failure cases.

A. Streetlight Map Generation

We employ an image-based object detection method [15] to filter luminous streetlight points from LiDAR maps or measurements for generating the pre-built streetlight map. For each LiDAR scan, points whose projections fall into the detected streetlight boxes are labeled as “streetlight”. The original LiDAR map is generated by the modified FAST-LIO2¹ [16] augmented with a fast and accurate loop closure module [17] and only the map points with “streetlight” labels are retained. The streetlight point map is further filtered and clustered by DBSCAN [18] for grouping 3D points into different streetlight instances. Finally the neat 3D streetlight map can be generated as shown in the bottom of Fig. 1.

B. State Representation

The robot state includes the orientation $\mathbf{R}_b^w \in SO(3)$, position $\mathbf{p}_b^w \in \mathbb{R}^3$, velocity ${}^w\mathbf{v}_b \in \mathbb{R}^3$, as well as the IMU accelerometer bias $\mathbf{b}_a \in \mathbb{R}^3$ and gyroscope bias $\mathbf{b}_\omega \in \mathbb{R}^3$.

The orientation \mathbf{R}_b^w , position \mathbf{p}_b^w and velocity ${}^w\mathbf{v}_b$ evolve on the matrix Lie group $SE_2(3)$ [19]. The element $\mathbf{X}_t \in SE_2(3)$ of time t is represented in the 5×5 matrix form:

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{R}_{b_t}^w & {}^w\mathbf{v}_{b_t} & \mathbf{p}_{b_t}^w \\ \mathbf{0}_{1 \times 3} & 1 & 0 \\ \mathbf{0}_{1 \times 3} & 0 & 1 \end{bmatrix} \quad (1)$$

We use the right invariant error $\boldsymbol{\eta}_t^r$ [20] with elements lying in Lie algebra $\mathfrak{se}_2(3)$ which is formulated by first-order approximation as below:

$$\boldsymbol{\eta}_t^r = \hat{\mathbf{X}}_t \mathbf{X}_t^{-1} \approx \mathbf{I} + \boldsymbol{\xi}_t^\wedge = \begin{bmatrix} \mathbf{I} + (\boldsymbol{\xi}_{R_t})_\times & \boldsymbol{\xi}_{v_t} & \boldsymbol{\xi}_{p_t} \\ \mathbf{0}_{1 \times 3} & 1 & 0 \\ \mathbf{0}_{1 \times 3} & 0 & 1 \end{bmatrix} \quad (2)$$

where $(\cdot)^\wedge$ maps the Euclidean tangent space of Lie group to the corresponding Lie algebra space at the identity and $(\cdot)_\times$ denotes the 3×3 skew-symmetric matrix. \mathbf{I} is the group identity and \mathbf{I} is the 3×3 identity matrix. We then have:

$$\begin{aligned} \mathbf{I} + (\boldsymbol{\xi}_{R_t})_\times &\approx \hat{\mathbf{R}}_{b_t}^w \mathbf{R}_{b_t}^{w\top} \\ \boldsymbol{\xi}_{v_t} &= {}^w\hat{\mathbf{v}}_{b_t} - \hat{\mathbf{R}}_{b_t}^w \mathbf{R}_{b_t}^{w\top} {}^w\mathbf{v}_{b_t} \\ \boldsymbol{\xi}_{p_t} &= \hat{\mathbf{p}}_{b_t}^w - \hat{\mathbf{R}}_{b_t}^w \mathbf{R}_{b_t}^{w\top} \mathbf{p}_{b_t}^w \end{aligned} \quad (3)$$

It is unfortunate that IMU biases can not fit into a Lie group. Therefore, we adopt the design of state representation

¹<https://github.com/hku-mars/FAST-LIO>

in [21], [22] which treats the biases in a similar way as EKF. We use $\mathbf{b}_t = [\mathbf{b}_{\omega_t}^\top \ \mathbf{b}_{a_t}^\top]^\top \in \mathbb{R}^6$ to denote the bias states and the corresponding bias error is defined as

$$\xi_t = [\zeta_{\omega_t}^\top \ \zeta_{a_t}^\top]^\top = [(\hat{\mathbf{b}}_{\omega_t} - \mathbf{b}_{\omega_t})^\top \ (\hat{\mathbf{b}}_{a_t} - \mathbf{b}_{a_t})^\top]^\top \quad (4)$$

The system state can be propagated by IMU kinematic propagation model [19] and the observation model for the correction step is given as

$$\mathbf{y}_t = h(\mathbf{X}_t) + \mathbf{v}_t \quad (5)$$

where \mathbf{y}_t and \mathbf{v}_t are the measurement vector and Gaussian noise vector, respectively. $h(\cdot)$ represents the observation model. To update the state with a new measurement, we use the form $\mathbf{z}_t = \mathbf{y}_t - h(\hat{\mathbf{X}}_t)$ as measurement error. Substituting (5) into the error function and expanding it in a first-order Taylor series approximation, we can get:

$$\mathbf{z}_t = -\mathbf{H}_t \xi_t + \mathbf{v}_t \quad (6)$$

where \mathbf{H}_t is the observation Jacobian matrix. From the linearized observation model (6), we can derive the Kalman gain \mathbf{K}_t . Meanwhile, the state estimate and the covariance matrix \mathbf{P}_t can be updated to posterior values.

C. IMU Propagation Model

The IMU measurements corrupted by the accelerometer bias, gyroscope bias, and additive Gaussian white noises are

$$\tilde{\omega}_t = \omega_t + \mathbf{b}_{\omega_t} + \mathbf{n}_{\omega_t}, \quad \tilde{a}_t = \mathbf{a}_t + \mathbf{b}_{a_t} + \mathbf{n}_{a_t} \quad (7)$$

where $\tilde{\omega}_t$ and \tilde{a}_t represent measured measurements of angular velocity and specific force in the IMU frame (i.e. the body frame b). $\mathbf{n}_{\omega_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma_{\omega})$ and $\mathbf{n}_{a_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma_a)$ are the noises. The state dynamics model can be derived with the IMU propagation model [22] in continuous-time as

$$\begin{aligned} \dot{\mathbf{R}}_{b_t}^w &= \mathbf{R}_{b_t}^w (\tilde{\omega}_t - \mathbf{b}_{\omega_t} - \mathbf{n}_{\omega_t}) \times \\ {}^w \dot{\mathbf{v}}_{b_t} &= \mathbf{R}_{b_t}^w (\tilde{a}_t - \mathbf{b}_{a_t} - \mathbf{n}_{a_t}) + \mathbf{g}^w \\ \dot{\mathbf{p}}_{b_t}^w &= {}^w \mathbf{v}_{b_t} \\ \dot{\mathbf{b}}_{\omega_t} &= \mathbf{n}_{b\omega_t}, \quad \dot{\mathbf{b}}_{a_t} = \mathbf{n}_{ba_t} \end{aligned} \quad (8)$$

The biases are modeled as random walks, and thus the derivatives of biases follow zero-mean Gaussian distributions, i.e., $\mathbf{n}_{b\omega_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma_{b\omega})$, $\mathbf{n}_{ba_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma_{ba})$.

For the states fitting into matrix Lie groups, by carrying out chain rules and the first-order approximation: $\eta_t^T = \exp(\xi_t) \approx \mathbf{I} + \xi_t^\wedge$, the linearized model of right invariant error can be derived as

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} \xi_{R_t} \\ \xi_{v_t} \\ \xi_{p_t} \\ \zeta_{\omega_t} \\ \zeta_{a_t} \end{bmatrix} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & -\hat{\mathbf{R}}_{b_t}^w & \mathbf{0} \\ (\mathbf{g}^w) \times & \mathbf{0} & \mathbf{0} & -({}^w \hat{\mathbf{v}}_{b_t}) \times \hat{\mathbf{R}}_{b_t}^w & -\hat{\mathbf{R}}_{b_t}^w \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & -(\hat{\mathbf{p}}_{b_t}^w) \times \hat{\mathbf{R}}_{b_t}^w & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \xi_{R_t} \\ \xi_{v_t} \\ \xi_{p_t} \\ \zeta_{\omega_t} \\ \zeta_{a_t} \end{bmatrix} + \\ & \begin{bmatrix} \hat{\mathbf{R}}_{b_t}^w & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ ({}^w \hat{\mathbf{v}}_{b_t}) \times \hat{\mathbf{R}}_{b_t}^w & \hat{\mathbf{R}}_{b_t}^w & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (\hat{\mathbf{p}}_{b_t}^w) \times \hat{\mathbf{R}}_{b_t}^w & \mathbf{0} & \hat{\mathbf{R}}_{b_t}^w & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{n}_{\omega_t} \\ \mathbf{n}_{a_t} \\ \mathbf{0}_{3 \times 1} \\ \mathbf{n}_{b\omega_t} \\ \mathbf{n}_{ba_t} \end{bmatrix} \triangleq \mathbf{A}_t \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \mathbf{Ad}_{\hat{\mathbf{X}}_t, \hat{\mathbf{b}}_t} \mathbf{n}_t \end{aligned} \quad (9)$$

where $\mathbf{Ad}_{\hat{\mathbf{X}}_t, \hat{\mathbf{b}}_t} \triangleq \text{block_diag}(\mathbf{Ad}_{\hat{\mathbf{X}}_t}, \mathbf{I}_{6 \times 6})$ and $\mathbf{Ad}_{\hat{\mathbf{X}}_t}$ is the adjoint matrix of $\hat{\mathbf{X}}_t$ [19]. $\mathbf{0}$ is the 3×3 zero matrix. The supplementary material [23] gives the detailed derivation. With the above continuous-time propagation model and (8),

the discretized system dynamics can be derived using Euler integration [22]:

$$\begin{aligned} \hat{\mathbf{R}}_{b_t+\Delta t}^{w-} &= \hat{\mathbf{R}}_{b_t}^{w+} \exp((\tilde{\omega}_t - \hat{\mathbf{b}}_{\omega_t}^+) \Delta t) \\ {}^w \hat{\mathbf{v}}_{b_t+\Delta t}^w &= {}^w \hat{\mathbf{v}}_{b_t}^+ + (\hat{\mathbf{R}}_{b_t}^{w+} (\tilde{a}_t - \hat{\mathbf{b}}_{a_t}^+) + \mathbf{g}^w) \Delta t \\ \hat{\mathbf{p}}_{b_t+\Delta t}^{w-} &= \hat{\mathbf{p}}_{b_t}^{w+} + {}^w \hat{\mathbf{v}}_{b_t}^+ \Delta t + \frac{1}{2} (\hat{\mathbf{R}}_{b_t}^{w+} (\tilde{a}_t - \hat{\mathbf{b}}_{a_t}^+) + \mathbf{g}^w) \Delta t^2 \\ \hat{\mathbf{b}}_{\omega_t+\Delta t}^- &= \hat{\mathbf{b}}_{\omega_t}^+, \quad \hat{\mathbf{b}}_{a_t+\Delta t}^- = \hat{\mathbf{b}}_{a_t}^+ \end{aligned} \quad (10)$$

where Δt is the time step between two consecutive IMU samples. $(\cdot)^-$ and $(\cdot)^+$ represent the predicted and the updated state. The IMU measurements are assumed to be constant between adjacent samplings. The covariance is computed according to the Riccati equation [24]:

$$\hat{\mathbf{P}}_{t+\Delta t}^- = \Phi_t \hat{\mathbf{P}}_t^+ \Phi_t^\top + \Phi_t \mathbf{A} d_{\hat{\mathbf{X}}_t, \hat{\mathbf{b}}_t} \mathbf{Q}_t \mathbf{A} d_{\hat{\mathbf{X}}_t, \hat{\mathbf{b}}_t}^\top \Phi_t^\top \Delta t \quad (11)$$

where $\Phi_t = \exp_m(\mathbf{A}_t \Delta t)$ is the state transition matrix, $\exp_m(\cdot)$ is the exponential map of matrices. $\hat{\mathbf{P}}_t^+$ is the posterior covariance of ξ_t at time t . $\mathbf{Q}_t = \text{block_diag}(\Sigma_{\omega}, \Sigma_a, \mathbf{0}, \Sigma_{b\omega}, \Sigma_{ba})$ is the covariance of \mathbf{n}_t .

D. Odometer Observation Model

The wheel odometer with encoders can provide linear velocity measurements for wheel robots. The measured velocity is also assumed to be corrupted by Gaussian white noises

$${}^o \tilde{\mathbf{v}}_{o_t} = {}^o \mathbf{v}_{o_t} + \mathbf{n}'_{o_t} \quad (12)$$

where $\mathbf{n}'_{o_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma'_{o_t})$. The velocity expressed in the IMU frame ${}^b \tilde{\mathbf{v}}_{b_t}$ and its covariance can be transformed through the calibrated extrinsic parameters $(\mathbf{R}_o^b, \mathbf{t}_o^b)$ between the odometer and body (IMU) frame

$${}^b \tilde{\mathbf{v}}_{b_t} = \mathbf{R}_o^b {}^o \tilde{\mathbf{v}}_{o_t} + (\mathbf{t}_o^b) \times (\tilde{\omega}_t - \mathbf{b}_{\omega_t}) \quad (13)$$

$$\Sigma_o = \mathbf{R}_o^b \Sigma'_o \mathbf{R}_o^{b\top} + (\mathbf{t}_o^b) \times (\Sigma_{\omega} + \Sigma_{b\omega}) (\mathbf{t}_o^b) \times^\top \quad (14)$$

In our case, the IMU and odometer are assumed to reside in the same location with only \mathbf{R}_o^b being considered. The observation model under the right invariant form is

$$\mathbf{y}_t = {}^b \tilde{\mathbf{v}}_{b_t} = h(\mathbf{X}_t) + \mathbf{n}_{o_t}, \quad h(\mathbf{X}_t) = \mathbf{R}_{b_t}^{w\top} {}^w \mathbf{v}_{b_t} \quad (15)$$

where $\mathbf{n}_{o_t} \sim \mathcal{N}(\mathbf{0}_{3 \times 1}, \Sigma_o)$. As in (6), the measurement error can be derived as

$$\mathbf{z}_{o_t} = -[\mathbf{0} \ \mathbf{I} \ \mathbf{0} \ \mathbf{0} \ \mathbf{0}] \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \hat{\mathbf{R}}_{b_t}^w \mathbf{n}_{o_t} \triangleq -\mathbf{H}_{o_t} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \hat{\mathbf{R}}_{b_t}^w \mathbf{n}_{o_t} \quad (16)$$

Using the Kalman filtering theory [20], the updated states and covariance can be obtained.

E. Camera Observation Model

1) *Data Association:* Assuming there are n_t detected streetlights and m_t 3D streetlight clusters within a certain range of th_{Lr1} from the robot pose estimate $\hat{\mathbf{R}}_{b_t}^w, \hat{\mathbf{p}}_{b_t}^w$. The streetlight observations are represented as $\{B_{t_i} = (\tilde{\mathbf{p}}_{t_i}, \tilde{\mathbf{l}}_{t_i})\}_{i=1}^{n_t}$ where $\tilde{\mathbf{p}}_{t_i} = [\tilde{u}_{t_i} \ \tilde{v}_{t_i}]^\top$ is the image coordinate of the center of object detection bounding boxes. $\tilde{\mathbf{l}}_{t_i} = [\tilde{l}_{t_i} \ \tilde{w}_{t_i}]^\top$ contains the width and length of box. The streetlight clusters in the map are represented as $\{L_j = (\tilde{\mathbf{C}}_j^w, {}^1 \tilde{\mathbf{P}}_j^w, \dots, {}^J \tilde{\mathbf{P}}_j^w), (-1)\}_{j=1}^{m_t+1}$ where $\tilde{\mathbf{C}}_j^w = \frac{1}{j} \sum_{j=1}^j {}^{jj} \tilde{\mathbf{P}}_j^w$ is the geometric center and ${}^{jj} \tilde{\mathbf{P}}_j^w \in \mathbb{R}^3$ denotes the 3D coordinate of each point. Since there may be false detections of streetlights, the set of streetlight clusters is extended with element (-1) . We score all combinations of matches between 2D and 3D data and the one with the

highest score is determined as the correct data association. The score is calculated based on two kinds of match errors.

Reprojection Error. Given a prior pose, an incorrect match generally results in a large error between the projected center of the 3D streetlight cluster and the box center. We model the residual vector $\mathbf{r}_{t_{proj}}$ as Gaussian distribution. Then for detected box B_{t_i} and streetlight cluster L_j , the reprojection error-based score $s_{t_{proj},ij}$ is calculated as

$$s_{t_{proj},ij} = \mathcal{N}(\mathbf{r}_{t_{proj},ij} | \mathbf{0}_{3 \times 1}, \Sigma_{t_{proj},ij})$$

$$\mathbf{r}_{t_{proj},ij} = \tilde{\mathbf{p}}_{t_i} - \pi_c(\mathbf{R}_b^c \hat{\mathbf{R}}_{b_t}^{w+} (\tilde{\mathbf{C}}_j^w - \hat{\mathbf{p}}_{b_t}^w) + \mathbf{t}_b^c) \quad (17)$$

where $\mathcal{N}(\cdot | \mathbf{0}_{3 \times 1}, \Sigma_{t_{proj},ij})$ represents the Gaussian probability density function with mean $\mathbf{0}_{3 \times 1}$ and covariance $\Sigma_{t_{proj},ij}$ in the 3D Euclidean space. \mathbf{R}_b^c and \mathbf{t}_b^c constitute the extrinsics between IMU and camera. $\pi_c(\cdot)$ is the perspective projection function of the camera.

Angle Error. We also introduce the score based on angle error θ_{ij} between the back-projection ray of the center of box B_{t_i} and the line connecting the center of streetlight cluster L_j and the optical center of the camera, which is defined as

$$s_{t_{ang},ij} = \mathcal{N}(r_{t_{ang},ij} | 0, \sigma_{t_{ang},ij}^2)$$

$$r_{t_{ang},ij} = 1 - \cos(\theta_{ij}) = 1 - \frac{\tilde{\mathbf{P}}_{t_i}^\top \tilde{\mathbf{C}}_j^c}{\|\tilde{\mathbf{P}}_{t_i}\|_2 \cdot \|\tilde{\mathbf{C}}_j^c\|_2} \quad (18)$$

where $\tilde{\mathbf{P}}_{t_i} = \pi_c^{-1}(\tilde{\mathbf{p}}_{t_i})$ is the back-projection of box center. The derivations of $\Sigma_{t_{proj},ij}$ and $\sigma_{t_{ang},ij}^2$ are provided in the supplementary material.

Then the total score of each match is formulated as

$$s_{ij} = \omega s_{t_{proj},ij} + (1 - \omega) s_{t_{ang},ij} \quad (19)$$

where $0 \leq \omega \leq 1$ is the preset weight. For the streetlight i without any correspondence, the score is calculated as

$$s_{ij} = 1 - \sum_{m=1}^{m_t} s_{im}, \quad j > m_t \quad (20)$$

We use d_{ij} to indicate whether the i -th streetlight observation matches the j -th 3D streetlight cluster. $\mathcal{D}_t = \{(B_{t_i}, L_j) | d_{ij} = 1\}_{i=1}^{n_t}$ denotes the possible match combinations. Note that all streetlight observations have a one-to-one mapping with 3D clusters. We expand the set of clusters to $m_t + n_t$ with n_t elements representing no correspondence of 2D observations so that this one-to-one mapping feature can be maintained when observation B_i is mapped to (-1) . To obtain the best combination \mathcal{D}_t^* , the optimization problem is formulated as the classical assignment problem

$$\mathcal{D}_t^* = \arg \max_{\mathcal{D}_t} \sum_{i=1}^{n_t} \sum_{j=1}^{m_t+n_t} d_{ij} s_{ij}, \quad \text{s.t.} \quad \sum_{i=1}^{n_t} d_{ij} \leq 1 \quad \sum_{j=1}^{m_t+n_t} d_{ij} = 1 \quad (21)$$

The optimization problem can be fast solved (~ 0.2 ms per frame) by Hungarian algorithm [25].

After obtaining the 2D-3D matches, the camera observation model can be derived. Only matches of \mathcal{D}_t^* without $L_j = (-1)$ are selected. For the i -th match (B_{t_i}, L_j) , the observation model is formulated as

$$\mathbf{y}_{t_i} = \pi_c^{-1}(\tilde{\mathbf{p}}_{t_i}) = h_i(\mathbf{X}_t) + \mathbf{n}_c$$

$$h_i(\mathbf{X}_t) = [\mathbf{R}_b^c \hat{\mathbf{R}}_{b_t}^{w+} (\tilde{\mathbf{C}}_{t_j}^w - \mathbf{p}_{b_t}^w) + \mathbf{t}_b^c] / Z_{t_j}^c \quad (22)$$

where $\mathbf{n}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^{-1} \Sigma_c \mathbf{K}^{-\top})$ and \mathbf{K} is the camera intrinsics matrix. Σ_c is a 3×3 diagonal matrix with 0

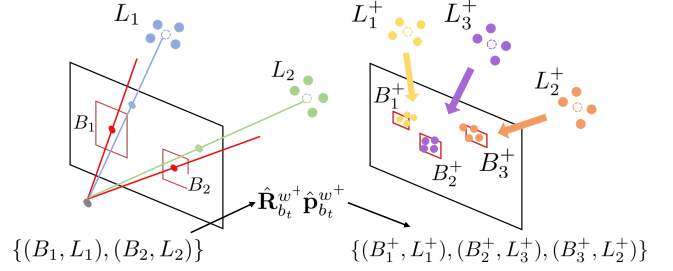


Fig. 3: An example of match extension. The module determines more matches according to the positional relationship between the projected locations of streetlight clusters and the boxes detected by the binary segmentation method.

as the last element. According to (6), by defining $\hat{\Psi}_{t_i} \triangleq \hat{\mathbf{R}}_{b_t}^{w+} (\tilde{\mathbf{C}}_{t_j}^w - \hat{\mathbf{p}}_{b_t}^w)$, the observation matrix can be derived as

$$\mathbf{z}_{c_{t_i}} = -\mathbf{H}_i(\hat{\Psi}_{t_i}) \hat{\mathbf{R}}_{b_t}^{w+ \top} [(\tilde{\mathbf{C}}_{t_j}^w \times \mathbf{0} - \mathbf{I} \mathbf{0} \mathbf{0})] \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \mathbf{n}_c$$

$$\triangleq -\mathbf{H}_{c_{t_i}} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \mathbf{n}_c \quad (23)$$

where $\mathbf{H}_i(\hat{\Psi}_{t_i}) \triangleq \frac{\partial h_i(\Psi_{t_i})}{\partial \Psi_{t_i}} \Big|_{\Psi_{t_i} = \hat{\Psi}_{t_i}}$. The detailed derivation is given in the supplementary material. The measurement error is obtained as

$$\mathbf{z}_{c_t} = \begin{bmatrix} \mathbf{z}_{c_{t_1}} \\ \vdots \\ \mathbf{z}_{c_{t_{n_t}}} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{c_{t_1}} \\ \vdots \\ \mathbf{H}_{c_{t_{n_t}}} \end{bmatrix} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \begin{bmatrix} \mathbf{n}_c \\ \vdots \\ \mathbf{n}_c \end{bmatrix} \triangleq \mathbf{H}_{c_t} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} + \mathbf{\Pi} \mathbf{n}_c \quad (24)$$

where $\mathbf{\Pi} = [\mathbf{I} \ \dots \ \mathbf{I}]^\top \in \mathbb{R}^{3n_t \times 3}$.

2) *Match Extension:* Despite the streetlight detector can identify the streetlights precisely, small lights are easily ignored, leading to a shortage of viable streetlight matches. To remedy this problem, we introduce the binary segmentation-based streetlight detection method. Fig. 3 shows an example of match extension. Pixels with intensity larger than th_{int} are fed into the Suzuki85 [26] algorithm to generate multiple bounding boxes which are considered as streetlight observations. Then with the updated robot pose $\hat{\mathbf{R}}_{b_t}^{w+}$ and $\hat{\mathbf{p}}_{b_t}^{w+}$, all streetlight map clusters within a range of th_{LR2} and not having 2D observation matches are projected to the image. To take one streetlight cluster as an example, since the updated pose is relatively accurate, all boxes containing the projected points are considered as the candidate matches. For each box, we then calculate the ratio of the number of projected points falling into the box to the total number of points within the streetlight cluster. The box with the maximum ratio is determined as the match of the cluster. By traversing all streetlight clusters, multiple new matches are selected and the robot state can be updated again.

3) *Environmental Degeneration:* When the streetlights are in a straight line, as shown in Fig. 4, any camera pose lying on the blue dotted circle satisfies the geometric constraints, thus the pose can drift with time. After the robot experiences a period of degeneration, the matches of new streetlights that are not in a line with already observed streetlights are difficult to search as shown in Fig. 4. We further find the positional drift in z -axis and rotational drifts in roll and pitch are

more obvious (Section IV-C). The reason is that the odometer can provide 2D motion measurements thus constraining the 3-DoF (x , y , and yaw). This case results in the projected points of the newly observed streetlights typically being located directly above or below the detection box. Using this feature, a direct and effective way is proposed. When the degeneration ends (i.e., new streetlights not in a line with already observed ones are detected), each projected center of newly observed streetlight clusters is assigned a rectangle. If the rectangle has an intersection area with detected boxes, the box closest to the projected center is chosen as the match of the corresponding streetlight cluster. Then the pose can be corrected by adding streetlights not in a straight line. Note that the height of the rectangle is significantly larger than its width due to the significant drifts in z -axis, roll, and pitch.

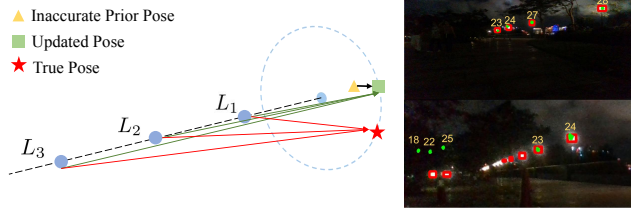


Fig. 4: Degeneration case. The left image shows that all positions on the blue circle satisfy the identical camera observation constraints. An inaccurate prior pose leads to deviation from the true pose. The right images show that the streetlight matches are difficult to search after the degeneration cases.

F. Tracking Recovery

In the cases without visible streetlights, the estimated trajectory could drift with time. The tracking recovery module is proposed to recover the pose estimate in the map. During tracking failures, brute-force matching is performed to traverse all match combinations. Assuming that there are n_{re} detected boxes and m_{re} streetlight clusters, each combination is assigned a punishment score calculated as:

$$s_{re} = \frac{1}{n_{re}} \sum_{i=1}^{n_{re}} \Delta p_i^{b_{re}+} + \gamma_1 \Delta t_{b_{re}}^{b_{re}+} + \gamma_2 \Delta \theta_{b_{re}}^{b_{re}+} + \gamma_3 \sum_{i=1}^{n_{re}} neg(i) \quad (25)$$

$$\Delta p_i^{b_{re}+} = \|\pi_c(\mathbf{R}_b^c \hat{\mathbf{R}}_{b_{re}}^{w+T} (\tilde{\mathbf{C}}_{re_j}^w - \hat{\mathbf{p}}_{b_{re}}^{w+}) + \mathbf{t}_b^c) - \tilde{\mathbf{p}}_{re_i}\|_2$$

$$\Delta t_{b_{re}}^{b_{re}+} = \|\hat{\mathbf{p}}_{b_{re},1,2}^{w-} - \hat{\mathbf{p}}_{b_{re},1,2}^{w+}\|_2$$

$$\Delta \theta_{b_{re}}^{b_{re}+} = yaw(\hat{\mathbf{R}}_{b_{re}}^{w-T} \hat{\mathbf{R}}_{b_{re}}^{w+})$$

where the updated pose ($\hat{\mathbf{R}}_{b_{re}}^{w+}, \hat{\mathbf{p}}_{b_{re}}^{w+}$) is calculated based on the combination. $neg(i) = 1$ when the i -th box has no correspondence, otherwise $neg(i) = 0$. $\gamma_1, \gamma_2, \gamma_3$ are utilized for balancing different terms. We only use the yaw angle of relative rotation ($yaw(\cdot)$) and the x, y components of relative translation ($(\cdot)_{1,2}$). Once there is a combination whose score is less than th_{RE_sc} and more than 2 matches are found, it is considered to successfully relocalize the robot. Subsequently, the updated pose of the combination with the minimum score is regarded as the recovered pose.

IV. EXPERIMENTS

A. Implementation Details

The robot is equipped with a front-facing RGB camera (30 Hz 1280×720 images), an odometer (10 Hz), and a

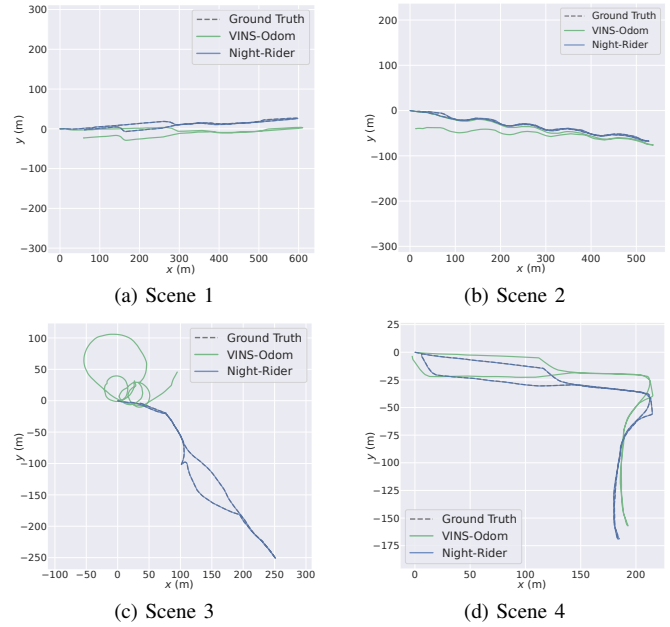


Fig. 5: Trajectories of ground truth (gray dashed line), VINS-Odom (green), and Night-Rider (blue) in four real scenes.

Livox 360° FOV LiDAR² (10 Hz) integrating an ICM40609 IMU (200 Hz). The camera's intrinsics and all extrinsics among sensors have been pre-calibrated. All experiments were performed on the Intel i7-1165G7 CPU. Note that only an IMU, an odometer, and a monocular camera are used for the proposed nocturnal vision-aided localization system.

We use Yolov5-Lite [15] as the streetlight detector for its real-time performance on CPU. The 2476 images labeled from different nighttime scenes are used to train the network. The relatively accurate initial pose is manually given and is optimized by the PnP solver [27] with streetlight matches.

B. Evaluation of Localization

To thoroughly evaluate the performance of the proposed system in nocturnal environments, we collect data in four different real scenes with multiple distribution types of streetlights. In the first scene (1149.2 m) and the fourth scene (723.8 m), the streetlights distribute sparsely on both sides of the road. Partial streetlights lie on only one side of the road in the second scene (1070.9 m). The third scene (777.2 m) contains densely distributed streetlights. All streetlight maps of these scenes are constructed using the method described in Section III-A. The offline optimized and precisely aligned trajectories from the loop closure augmented FAST-LIO2 using the LiDAR on board are treated as ground truths and EVO [28] is used for evaluation. Note that there are very few open-source research works targeting nocturnal scenarios. Thus we first compare Night-Rider with the state-of-the-art framework, extended VINS-Mono [3] with the odometer [29] (named VINS-Odom). As shown in Fig. 5, Night-Rider can precisely estimate trajectories with extremely low drifts in all four scenes while VINS-Odom drifts over time and even fails to track in the third scene. The reason is that the unstable illumination in nighttime scenes makes it difficult to extract

²<https://www.livoxtech.com/mid-360>

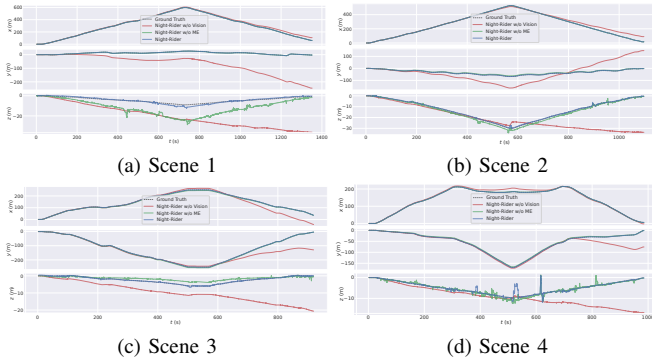


Fig. 6: Trajectories in x , y and z -axis of ground truth (gray dashed line), Night-Rider w/o Vision (red), Night-Rider w/o ME (green), and Night-Rider (blue) in four real scenes.

reliable feature points from highly bright or dark images. This instability results in drifts or tracking failures of VINS-Odom. Contrarily, Night-Rider is independent of feature points and exploits the stable streetlights for localization. Thus trajectories of Night-Rider align accurately with ground truths. Table. I depicts the evaluation results of absolute trajectory error (ATE [30]) for translation and rotation parts. It shows that Night-Rider achieves remarkably accurate and robust localization in all scenes ($<0.2\%$ relative error of total trajectory length) and totally outperforms VINS-Odom, which further validates the effectiveness of our system.

TABLE I: ATE [30] for translation and rotation of VINS-Odom, Night-Rider w/o Vision, Night-Rider w/o ME, and Night-Rider in four scenes. The best is shown in **bold**.

		VINS-Odom	Night-Rider w/o Vision	Night-Rider w/o ME	Night-Rider
Scene 1	Trans. (m)	20.770	108.707	8.617	1.543
	Rot. (deg)	2.959	15.553	4.071	2.307
Scene 2	Trans. (m)	24.200	71.678	3.398	0.589
	Rot. (deg)	5.521	21.444	2.640	1.831
Scene 3	Trans. (m)	221.422	57.864	4.072	0.339
	Rot. (deg)	109.834	19.841	2.554	1.674
Scene 4	Trans. (m)	12.992	23.100	2.345	0.880
	Rot. (deg)	5.445	13.088	2.872	3.176

We also perform ablation experiments to verify the proposed modules of our system. Three versions are used for comparison, namely the complete system (Night-Rider), Night-Rider without the match extension module (Night-Rider w/o ME), and Night-Rider without visual measurements (Night-Rider w/o Vision). Fig. 6 shows the position components of trajectories in four scenes. For Night-Rider w/o Vision, the estimated pose drifts over time, especially in the z -axis. With the proposed camera observation model in Night-Rider, the drift is significantly reduced. The trajectories of Night-Rider w/o ME deviate from the true poses, especially in the z -axis, since lots of distant streetlight observations are too small to be detected without the match extension module. More 2D-3D streetlight correspondences generated by the match extension module improve the accuracy and robustness of the whole system. The quantitative results in Table. I further verify the importance of the camera

observation model and the match extension module.

C. Tracking Lost and Degeneration Cases

Fig. 7 shows the estimated trajectory of scene 3 when the tracking lost or degeneration occurs. Here we display the z , roll, and pitch components of poses since the x , y , and yaw components accurately fit with the ground truth. The system without dealing with tracking lost or degeneration cases fail to estimate the robot poses. It can be observed that when tracking lost occurs (orange areas), the robot pose deviates significantly from the true value. Thanks to the proposed tracking recovery module, the pose can be immediately recovered to be close to the ground truth when streetlight observations arrive. The green area in Fig. 7 indicates that the robot experiences degenerated cases which result in pose drifting over time. When the robot observes streetlights without degeneration cases, the proposed strategy can effectively correct the drifted pose.

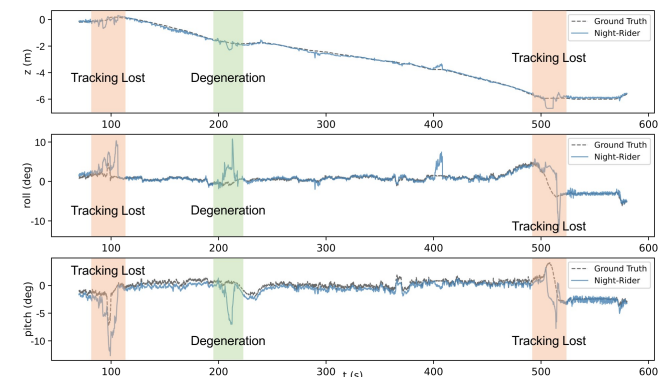


Fig. 7: Cases of tracking failure and degeneration. The proposed system can successfully recover the drifted pose.

V. CONCLUSION

In this paper, a nocturnal vision-aided localization system integrating camera, IMU, and odometer sensors is proposed for robot navigation at night. To address problems of nocturnal visual localization, we propose to leverage streetlight detections for nocturnal localization. We utilize the InEKF for state estimation and propose novel data association and match extension modules to fully exploit the information from streetlights. Experiments on real nighttime scenes validate that the proposed system achieves accurate and robust localization. Future work includes the integration with mainstream methods and other sensors for a more complete localization system which is adaptively available in both daytime and nighttime scenes.

ACKNOWLEDGEMENT

This work was supported in part by the Science and Technology Development Fund of Macao SAR under Grants 0046/2021/AGJ, 0067/2023/AFJ, 0123/2022/AFJ, and 0081/2022/A2.

REFERENCES

- [1] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [2] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [3] 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.
- [4] 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 Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [5] C. Zhang, H. Zhao, C. Wang, X. Tang, and M. Yang, “Cross-modal monocular localization in prior lidar maps utilizing semantic consistency,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4004–4010.
- [6] H. Wang, C. Xue, Y. Zhou, F. Wen, and H. Zhang, “Visual semantic localization based on hd map for autonomous vehicles in urban scenarios,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 255–11 261.
- [7] Z. Liao, J. Shi, X. Qi, X. Zhang, W. Wang, Y. He, X. Liu, and R. Wei, “Coarse-to-fine visual localization using semantic compact map,” in *2020 3rd International Conference on Control and Robots (ICCR)*. IEEE, 2020, pp. 30–37.
- [8] C. Guo, M. Lin, H. Guo, P. Liang, and E. Cheng, “Coarse-to-fine semantic localization with hd map for autonomous driving in structural scenes,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1146–1153.
- [9] H. Yu, W. Zhen, W. Yang, J. Zhang, and S. Scherer, “Monocular camera localization in prior lidar maps with 2d-3d line correspondences,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4588–4594.
- [10] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, “Monocular camera localization in 3d lidar maps,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1926–1931.
- [11] W.-C. Ma, I. Tartavull, I. A. Bârsan, S. Wang, M. Bai, G. Mattyus, N. Homayounfar, S. K. Lakshminanth, A. Pokrovsky, and R. Urtasun, “Exploiting sparse semantic hd maps for self-driving vehicle localization,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5304–5311.
- [12] M. Zhang, Y. Chen, and M. Li, “Vision-aided localization for ground robots,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2455–2461.
- [13] H. Gao, J. Guo, G. Wang, and Q. Zhang, “Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9913–9923.
- [14] P. Nelson, W. Churchill, I. Posner, and P. Newman, “From dusk till dawn: Localisation at night using artificial light sources,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5245–5252.
- [15] X. Chen and Z. Gong, “Yolov5-lite: Lighter, faster and easier to deploy,” *Accessed: Sep*, vol. 22, 2021.
- [16] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, “Fast-lid2: Fast direct lidar-inertial odometry,” *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.
- [17] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, “Lidar iris for loop-closure detection,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5769–5775.
- [18] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “Dbscan revisited, revisited: why and how you should (still) use dbscan,” *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.
- [19] R. Hartley, M. Ghaffari, R. M. Eustice, and J. W. Grizzle, “Contact-aided invariant extended kalman filtering for robot state estimation,” *The International Journal of Robotics Research*, vol. 39, no. 4, pp. 402–430, 2020.
- [20] A. Barrau and S. Bonnabel, “The invariant extended kalman filter as a stable observer,” *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1797–1812, 2016.
- [21] A. Barrau, “Non-linear state error based extended kalman filters with applications to navigation,” Ph.D. dissertation, Mines Paristech, 2015.
- [22] E. R. Potokar, K. Norman, and J. G. Mangelson, “Invariant extended kalman filtering for underwater navigation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5792–5799, 2021.
- [23] “Supplementary materials: Night-rider: Nocturnal vision-aided localization in streetlight maps using invariant extended kalman filtering,” https://github.com/IMRL/Night-Rider/blob/main/supplementary/supplementary_materials.pdf, 2023.
- [24] P. S. Maybeck, *Stochastic models, estimation, and control*. Academic press, 1982.
- [25] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [26] S. Suzuki *et al.*, “Topological structural analysis of digitized binary images by border following,” *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985.
- [27] V. Lepetit, F. Moreno-Noguer, and P. Fua, “Ep n p: An accurate o(n) solution to the p n p problem,” *International journal of computer vision*, vol. 81, pp. 155–166, 2009.
- [28] M. Grupp, “evo: Python package for the evaluation of odometry and slam,” <https://github.com/MichaelGrupp/evo>, 2017.
- [29] J. Liu, W. Gao, and Z. Hu, “Visual-inertial odometry tightly coupled with wheel encoder adopting robust initialization and online extrinsic calibration,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5391–5397.
- [30] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.