LiDAR-Aided Visual-Inertial Localization with Semantic Maps

Hao Li*, Liangliang Pan*, Ji Zhao

***Abstract*— Accurate and robust localization is an essential task for autonomous driving systems. In this paper, we propose a novel 3D LiDAR-aided visual-inertial localization method. Our method fully explores the complementarity of visual and LiDAR observations. On the one hand, the association between semantic features in images and a given semantic map provides constraints for the absolute pose. On the other hand, LiDAR odometry (LO) can provide an accurate and robust 6DOF relative pose. The Error State Kalman Filter (ESKF) framework is exploited to estimate the vehicle pose relative to the semantic map, which fuses the global constraints between the image and the semantic map, the relative pose from the LO, and the raw IMU data. The method achieves centimeter-level localization accuracy in a variety of challenging scenarios.**

We validate the robustness and accuracy of our method in real-world scenes over 50 km. The experimental results show that the proposed method is able to achieve an average lateral accuracy of 0.059 m and longitudinal accuracy of 0.158 m, which demonstrates the practicality of the proposed system in autonomous driving applications.

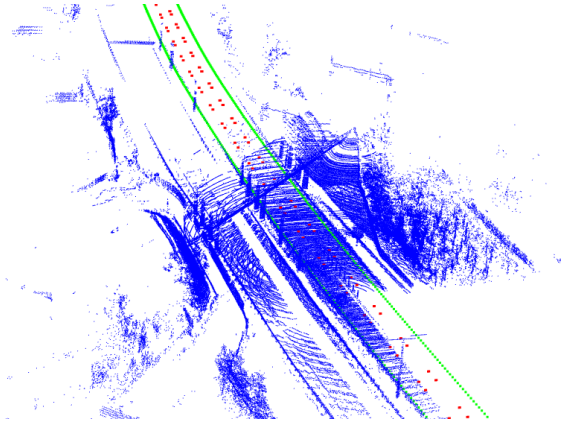


Fig. 1. Illustration of our online semantic map and local map. Our semantic HD map contains two types of lane points: solid points with green, and dashed points with red. The local map is represented by blue pointcloud, which is generated by LO online.

I. INTRODUCTION

Localization plays a critical role in autonomous driving systems. It is also essential for many other tasks, such as long-range visual perception and prediction. Cameras are one of the most common sensors used for autonomous driving due to their rich information and low cost. The proposed method in this paper also uses cameras as the primary sensor. Traditional visual features often have tracking lost due to lighting, weather, or view changes. Semantic features, especially lane lines, are common in urban and highways and are robust to the aforementioned changes. However, map-based localization methods are prone to pose drift due to environmental changes (e.g., road marks are repainted or occluded) or the lack of sufficient landmark observations in certain map areas.

Compared with the global information provided by visual features, LiDARs provide rich information for the local environment. Current LO methods can obtain accurate relative poses through the point cloud registration, which naturally complements the global visual observation constraints. Thus LO is useful for robust and accurate online localization and is not influenced by map changes and visual occlusion. While LO can obtain accurate relative poses in most scenarios, its pose stability will be affected in degraded scenarios (tunnels, sea-crossing bridges, etc.). The covariance of relative poses should be estimated to distinguish the pose stability, which effectively reflects the prior distribution of observations in state estimation problems. Motivated by the pose covariance estimation methods for ICP [1], [2], we extended them to general 3D ICP's covariance estimation (including point-to-point, point-to-edge, and point-to-plane types). In our method, LiDAR pose covariance is used to weight relative pose constraints for drift-aware pose fusion. By analyzing the eigenvalue of the information matrix (inverse matrix of covariance), we can also explicitly detect the under-constrained direction in degenerate cases.

The main contributions in this work are summarized as follows: (1) To cope with occlusion and scene degradation environments, we employ a tightly-coupled ESKF to fuse global semantic constraints and local odometry measurements. (2) We exploit a hybrid visual measurements model that fused both implicit distance loss and explicit reprojection loss to enhance accuracy and robustness. (3) A drift-aware LO with closed pose covariance formulation is implemented. It improves system robustness when applied to online fusion localization. (4) We conduct various large-scale road tests to validate the effectiveness of the developed method.

The paper is organized as follows. In Sec. II, we introduce the related works. In Sec. III, the tightly coupled ESKF system is presented with several sub-modules. Finally, the experimental results of our method are analyzed and compared in Sec. IV.

II. RELATED WORK

A. Geometrical Feature-based Methods

Traditional visual-inertial odometry (VO/VIO) methods [3]–[5] tend to use geometrical features, such as points,

* The first two authors contributed equally to this work. J. Zhao is the corresponding authors. All authors are with TuSimple, Beijing, China. Emails: hao.li@tusimple.ai, {ll.pan931204, zhaoji84}@gmail.com.

lines, and planes for feature representation and matching. Leutenegger et al. [3] extract Harris corner points and match corners using BRIEF descriptor. Furthermore, Pumarola et al. [4] explore to handle both point and line features. Salas-Moreno et al. Pan et al. [5] estimate the camera motion using both points and planes. However, the VO/VIO suffers from accumulated drift due to the lack of absolute observation constraints.

To solve the problem of accumulated drift in long-distance localization tasks, coupling GNSS with VO tends to be a promising solution. These methods [6]–[8] fuse GNSS measurement with visual features to estimate camera pose with the loose-coupled or tight-coupled method. However, these methods can only achieve decimeter-level or even meter-level precision. Besides, GNSS signals are easily blocked by buildings and tunnels for automatic driving tasks.

Some researchers also attempted to extend the VO system by adding a prior map to perform long-distance localization tasks. Campos et al. [9] and Se et al. [10] build the visual feature map with ORB and SIFT features respectively, and estimate the camera pose by matching image features with the map. However, geometrical features are suffered from illumination, perspective, and time changes in the long term.

B. Semantic Feature-based Methods

Compared to geometrical features, semantic features are more robust against time and illumination changes. Semantic feature-based methods build a prior map of road features, such as poles, lanes, and traffic signs, to perform localization in large-scale environments. Schreiber et al. [11] build a sparse map with lane lines. Ranganathan et al. [12] perform visual localization utilizing a sparser prior map with road markings. Zhang et al. [13] exploit both poles and lane features matching with a pre-build semantic map. Despite the success in practice, these methods are sensitive to feature missing, such as occlusion or lane repainting.

Multisensor fusion is a promising way to fix the problem. Ma et al. [14] fuse semantic features, GPS, and vehicle odometry (IMU & Wheel Encoder) in a particle filter framework. Huayou et al. [15] fuse semantic measurements and vehicle odometry in an optimization framework. However, the wheel odometer is sensitive to changes in tire pressure, wheel spin, and rigidity with the vehicle body. Furthermore, none of these methods model the confidences about external relative constraints.

This paper fuses LO with visual semantic and inertial information in a tightly-coupled ESKF framework. It has the following three key components. First, keypoint constraints and distance transform constraints from lane features are fused, constructing a global visual observation model. These two kinds of constraints provide complementary lateral and longitudinal constraints. Second, fusing the relative observation model from LO makes our system robust to some visual failure cases. Third, an accurate closed-form covariance estimation method is adopted to adjust the weight of LO constraints in real-time to enhance the accuracy and robustness of visual localization.

TABLE I
SOME IMPORTANT NOTATIONS

Symbols	Meaning
$\mathscr{F}$	The coordinate system. $\mathscr{F}_G$, $\mathscr{F}_B$, and $\mathscr{F}_C$ represent the global ENU frame, IMU frame and camera frame, respectively.
$\mathbf{x}$, $\bar{\mathbf{x}}$, $\hat{\mathbf{x}}$	The ground-truth, updated, and estimated global state. $\hat{\mathbf{x}} = \bar{\mathbf{x}}$ after filter update.
$\delta\mathbf{x}$	The error state between the ground-truth global state $\mathbf{x}$ and its estimation $\hat{\mathbf{x}}$.
$\Delta\mathbf{x}$, $\Delta\hat{\mathbf{x}}$	The groud-truth and estimated local state.
$\widetilde{\mathbf{x}}$	The error state between the ground-truth local state $\Delta\mathbf{x}$ and the estimated local state $\Delta\hat{\mathbf{x}}$.
$\mathbf{P}$	Covariance matrix. $\bar{\mathbf{P}}$ and $\hat{\mathbf{P}}$ represent the covariance matrix of $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$, respectively.
d_i, s_j	The i-th associated map point of dashed lanes and the j-th associated map point of solid lanes in an image frame.
${}^a\mathbf{p}_{d_i}, {}^a\mathbf{p}_{s_j}$	The position of map point d_i and s_j in $\mathscr{F}_a$.
$\mathbf{z}_{d_i}, \mathbf{z}_{s_j}$	The residual of visual observation for the map point d_i and s_j.
$\mathbf{z}_p$, $\mathbf{z}_r$	The residual of pose measurement for the position and rotation.
${}^a\mathbf{T}_b$	${}^a\mathbf{T}_b \in \mathrm{SE}(3)$ represents the transformation from $\mathscr{F}_b$ to $\mathscr{F}_a$.
${}^a\mathbf{R}_b$	${}^a\mathbf{R}_b \in \mathrm{SO}(3)$ represents the rotation matrix from $\mathscr{F}_b$ to $\mathscr{F}_a$.
${}^a\mathbf{p}_b$	${}^a\mathbf{p}_b \in \mathbb{R}^3$ represents the position of $\mathscr{F}_b$ in $\mathscr{F}_a$.
${}^a\mathbf{v}_b$	The velocity of $\mathscr{F}_b$ relative to $\mathscr{F}_a$.
$\mathbf{b_a}$, $\mathbf{b}_\omega$	The bias of accelerometer and gyroscope.

III. METHODOLOGY

A. System Overview

The notations in this paper are summarized in Table I. The overview of our method is shown in Figure 2. The camera and LiDAR are synchronized by hardware at 10Hz. The IMU built in LiDAR gathers movement data up to 200Hz. The images are fed into the Visual Measurement Module to extract dashed lanes and solid lanes. The extracted features will be associated with the semantic map. The point frames are fed into the LiDAR Odometry Module to estimate the relative pose between two frames. The Pose Measurement Module uses the updated state of last time and relative pose to infer the pose of the current time. The IMU measurements are fed into the Visual Measurement Module and Pose Measurement Module to calculate residuals for state estimation at 10Hz.

B. System Description

1) Coordinate Definition:

$\mathscr{F}_G$ denotes global ENU (East-North-Up) coordinate. The x, y, and z axes point to the East, North, and Up respectively, and its origin is a fixed point. $\mathscr{F}_{C_k}$, $\mathscr{F}_{B_k}$, and $\mathscr{F}_{L_k}$ denote camera, IMU, and LiDAR coordinates at time k respectively. Camera, LiDAR, and IMU are rigidly attached together with the extrinsic transformation ${}^B\mathbf{T}_C \doteq \left({}^B\mathbf{R}_C, {}^B\mathbf{p}_C\right)$ from camera frame $\mathscr{F}_C$ to IMU frame $\mathscr{F}_B$ and the extrinsic transformation ${}^B\mathbf{T}_L \doteq \left({}^B\mathbf{R}_L, {}^B\mathbf{p}_L\right)$ from LiDAR frame $\mathscr{F}_L$ to IMU frame $\mathscr{F}_B$. The extrinsic parameters are calibrated offline.

2) State Definition:

We take the IMU frame as the body frame and estimate the pose of the body frame relative to the ENU frame. Denoting $\mathbf{x}_k$ as the global state at time k:

$$\mathbf{x}_k \doteq \left[{}^G\mathbf{p}_{B_k}^T, {}^G\mathbf{v}_{B_k}^T, {}^G\mathbf{R}_{B_k}^T, \mathbf{b}_\mathbf{a}^T, \mathbf{b}_\omega^T\right]^T \tag{1}$$

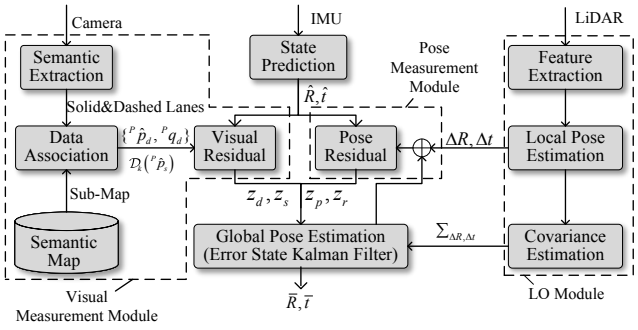
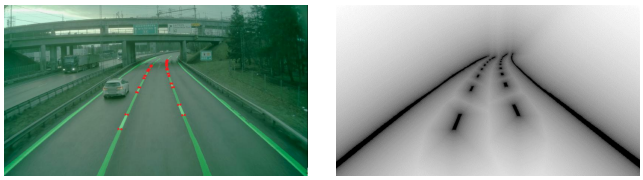


Fig. 2. Pipeline of the proposed method.



(a) Lane detection. (b) distance transform (DT) image.

Fig. 3. Visual observation. **Left**: raw image with keypoints (red dots) and lane detection (green masks). **Right**: DT image. Each pixel represents the euclidean distance from current pixel to the nearest lane region. Brighter means longer distance.

where ${}^{G}\mathbf{p}_{B_k}$, ${}^{G}\mathbf{v}_{B_k}$, and ${}^{G}\mathbf{R}_{B_k}$ denote the position, velocity, and attitude of IMU coordinate $\mathscr{F}_{B_k}$ in the global ENU coordinate $\mathscr{F}_G$ at time k. $\mathbf{b_a}$ and $\mathbf{b}_{\boldsymbol{\omega}}$ denote the bias of accelerometer and gyroscope at time k respectively.

We use error state $\delta\mathbf{x}_k$ to indicate the error between the ground-truth state $\mathbf{x}_k$ and the state estimation $\hat{\mathbf{x}}_k$:

$$\delta\mathbf{x}_k = \left[\delta\mathbf{p}_k^T, \delta\mathbf{v}_k^T, \delta\boldsymbol{\theta}_k^T, \delta\mathbf{b}_{\mathbf{a}}^T, \delta\mathbf{b}_{\boldsymbol{\omega}}^T\right]^T \tag{2}$$

where $\delta\boldsymbol{\theta} = \log\left({}^{G}\overline{\mathbf{R}}_B^T {}^{G}\mathbf{R}_B\right) \in \mathbb{R}^3$ is the attitude error vector.

The LO module output relative poses between two adjacent point cloud frames. Let $\Delta\mathbf{x}_{L_k}$ denote the local state:

$$\Delta\mathbf{x}_{L_k} \doteq \left[{}^{L_k}\mathbf{p}_{L_{k+1}}^T, {}^{L_k}\mathbf{R}_{L_{k+1}}^T\right]^T \tag{3}$$

where ${}^{L_k}\mathbf{p}_{L_{k+1}}$ and ${}^{L_k}\mathbf{R}_{L_{k+1}}$ denote the position and attitude of $\mathscr{F}_{L_{k+1}}$ in $\mathscr{F}_{L_k}$ respectively.

3) State Propagation:

Once receiving the IMU measurement, we can perform state propagation based on the discrete inertial kinematic model $\mathbf{f}(\cdot)$ [16] under the assumption of zero noise:

$$\hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_i \boxplus (\Delta t \cdot \mathbf{f}(\hat{\mathbf{x}}_i, \mathbf{u}_i, \mathbf{0})) \tag{4}$$

where i is the index of IMU measurement. Δt is the IMU sampling period. $\mathbf{u}_i$ is the i-th IMU input, which is the stacked vector of linear acceleration measurement and angular velocity measurement. Operation $\boxplus$ maps the error state to the corresponding manifold [17].

We linearize Eq. (4) at the current estimated state and propagate the covariance in time:

$$\hat{\mathbf{P}}_{i+1} = \mathbf{F}_{\delta\mathbf{x}}\hat{\mathbf{P}}_i\mathbf{F}_{\delta\mathbf{x}}^T + \mathbf{F}_{\omega}\mathbf{Q}_i\mathbf{F}_{\omega}^T \tag{5}$$

where $\mathbf{Q}_i$, $\mathbf{F}_{\delta\mathbf{x}}$, and $\mathbf{F}_{\boldsymbol{\omega}}$ are the covariance of IMU measurement noises, state transition matrix, and noise transition matrix [16], respectively.

C. Visual Measurement Module

We adopt a CNN-based image processing method [18], [19] for lane segmentation and keypoint detection, as shown in Fig. 3(a). We parameterize corners of each dashed lane blob into four keypoints. The semantic lane mask is binarized into a binary image, in which pixels within the lane region are set to 0, and others are set to 1. Then we perform distance transform (DT) [20], [21] on a binary mask to get the DT image, as shown in Fig. 3(b). Each pixel value in a DT image denotes the euclidean distance from the current pixel to its closest lane point.

The semantic map contains dashed map points $\{d_i \in \mathbb{R}^3\}$ and sampled solid map points $\{s_j \in \mathbb{R}^3\}$. The coordinate of map points are aligned to the global ENU frame $\mathscr{F}_G$ and represented as ${}^{G}\mathbf{p}_{d_i}$ and ${}^{G}\mathbf{p}_{s_j}$ for $\{d_i\}$ and $\{s_j\}$, respectively. Fig. 4 shows an illustration of the DA method. Given an initial pose from IMU propagation in Eq. (4), these semantic cues detected from the image and their correspondences from the semantic feature map will be associated. In order to reduce poor effects caused by repetitive texture and bad lane curve fitting in complex road conditions, we apply an implicit data association (DA) method [22] for solid lane points to construct the distance loss, which produce obvious constraints in lateral directions. Morevever, for dashed keypoint observation constraints, the explicit method is adopted due to their sparsity and clarity with less possibility of mismatching. It associates the reprojected map point with the nearest detected corner point, which produce constraints both in lateral and longitudinal directions. With the estimated state $\hat{\mathbf{x}}_k$ from IMU propagation, we can project candidate 3D map points ${}^{G}\mathbf{p} = \{{}^{G}\mathbf{p}_{d_i}, {}^{G}\mathbf{p}_{s_i}\}$ in the semantic map onto an image and construct residuals by DA.

1) **Visual Keypoints Observation Model***:* The reprojection error is adopted for the keypoint of dashed lanes:

$$\mathbf{z}_{d_i} = {}^{P}\mathbf{q}_{d_i} - {}^{P}\hat{\mathbf{p}}_{d_i} \tag{6}$$

where ${}^{P}\hat{\mathbf{p}}_{d_i} \in \mathbb{R}^2$ is the projection of map point d_i with the estimated state $\hat{\mathbf{x}}_k$. ${}^{P}\mathbf{q}_{d_i} \in \mathbb{R}^2$ is the detected corner point associated with the projected point ${}^{P}\hat{\mathbf{p}}_{d_i}$. Considering the measurement noise, the visual keypoint observation model can be obtained:

$$\begin{aligned}\mathbf{0} &= \mathbf{h}_{d_i}\left(\mathbf{x}_k, \mathbf{n}_{d_i}\right) = {}^{P}\mathbf{q}_{d_i} - {}^{P}\mathbf{p}_{d_i} - \mathbf{n}_{d_i} \\ &\simeq \mathbf{h}_{d_i}(\hat{\mathbf{x}}_k, \mathbf{0}) + \mathbf{H}_{d_i}\delta\mathbf{x}_k - \mathbf{n}_{d_i} = \mathbf{z}_{d_i} + \mathbf{H}_{d_i}\delta\mathbf{x}_k - \mathbf{n}_{d_i}\end{aligned} \tag{7}$$

where $\mathbf{x}_k = \hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k$, $\mathbf{H}_{d_i}$ is the Jacobian matrix of $\mathbf{h}\left(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \mathbf{n}_{d_i}\right)$ with respect to $\delta\mathbf{x}_k$. ${}^{P}\mathbf{p}_{d_i}$ is the projection of corner point ${}^{G}\mathbf{p}_{d_i}$ in the semantic map with the ground-truth state $\mathbf{x}_k$. ${}^{P}\mathbf{q}_{d_i}$ is the corner point detection associated with the projected point ${}^{P}\mathbf{p}_{d_i}$. Gaussian white noise $\mathbf{n}_{d_i}$ of ${}^{P}\mathbf{q}_{d_i}$ originates from discrete pixels and unclear road surface. $\mathbf{n}_{d_i} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_{d_i}\right)$ in Eq. (7).

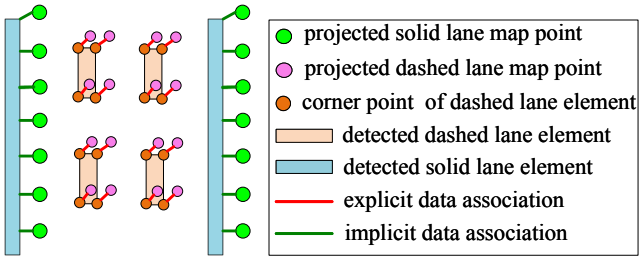


Fig. 4. Explicit DA method for keypoint of dashed lanes and implicit DA method for sampled solid lanes.

2) **Visual Distance Observation Model***:* Similarly, denote ${}^{P}\hat{\mathbf{p}}_{s_i}$ as the projection of the solid lane point ${}^{G}\hat{\mathbf{p}}_{s_i}$ in the semantic map with the estimated state $\hat{\mathbf{x}}_k$. DT distance is utilized to construct residual for solid lanes:

$$\mathbf{z}_{s_i} = \mathcal{D}_k\left({}^{P}\hat{\mathbf{p}}_{s_i}\right) \tag{8}$$

where $\mathcal{D}_k$ represents the look-up table distance loss from the k-th DT image. The visual distance observation model for each sampled solid lane point is:

$$\begin{aligned}\mathbf{0} &= \mathbf{h}_{s_j}\left(\mathbf{x}_k, \mathbf{n}_{s_j}\right) = \mathcal{D}_k\left({}^{P}\mathbf{p}_{s_j}\right) - \mathbf{n}_{s_j} \\ &\simeq \mathbf{h}_{s_j}\left(\hat{\mathbf{x}}_k, \mathbf{0}\right) + \mathbf{H}_{s_j}\delta\mathbf{x}_k - \mathbf{n}_{s_j} = \mathbf{z}_{s_j} + \mathbf{H}_{s_j}\delta\mathbf{x}_k - \mathbf{n}_{s_j}\end{aligned} \tag{9}$$

where $\mathbf{n}_{s_j} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_{s_j}\right)$ is the Gaussian white noise of the distance transforms of solid lanes. ${}^{P}\mathbf{p}_{s_j}$ is the projection of solid line map point ${}^{G}\mathbf{p}_{s_j}$ with the ground-truth state $\mathbf{x}_k$ and

$$\mathbf{H}_{s_j} = \left.\frac{\partial \mathbf{h}_{s_j}(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \mathbf{n}_{s_j})}{\partial \delta\mathbf{x}_k}\right|_{\delta\mathbf{x}_k=\mathbf{0}} = \left.\frac{\mathcal{D}_k({}^{P}\mathbf{p}_{s_j})}{{}^{P}\mathbf{p}_{s_j}} \frac{{}^{P}\mathbf{p}_{s_j}}{{}^{C}\mathbf{p}_{s_j}} \frac{{}^{C}\mathbf{p}_{s_j}}{\delta\mathbf{x}_k}\right|_{\delta\mathbf{x}_k=\mathbf{0}} \tag{10}$$

Here, ${}^{C}\mathbf{p}_{s_j}$ is a point where map point ${}^{G}\mathbf{p}_{s_j}$ is transformed from global ENU frame $\mathcal{F}_G$ to camera frame $\mathcal{F}_{C_k}$ with the ground-truth state $\mathbf{x}_k$.

D. LiDAR Odometry Module

Our LO Module adopts an optimization approach to estimate the local pose, which will be used in the Pose Measurement Module to provide relative pose constraints. Furthermore, we estimate covariance to analyze the uncertainty of local pose constraints in global localization.

1) Relative Pose Estimation:

In order to construct residuals for estimating local pose, we extract edge features $\mathbb{F}_e$ and planar features $\mathbb{F}_p$ according to the local smoothness as in [23]. The residual corresponding to the j-th feature point ${}^{L_{k+1}}\mathbf{p}_{l_j}$ collected in $\mathcal{F}_{L_{k+1}}$ is:

$$f_j(\Delta\mathbf{x}_{L_k}, {}^{L_{k+1}}\mathbf{p}_{l_j}) = \mathbf{G}_j\left({}^{L_k}\hat{\mathbf{p}}_{l_j} - {}^{L_k}\mathbf{q}_{l_j}\right) \tag{11}$$

where ${}^{L_k}\hat{\mathbf{p}}_{l_j}$ is the transformed feature point of ${}^{L_{k+1}}\mathbf{p}_{l_j}$ from $\mathcal{F}_{L_{k+1}}$ to $\mathcal{F}_{L_k}$ with the initial guess of local state $\Delta\hat{\mathbf{x}}_{L_k}$. The corresponding point ${}^{L_k}\mathbf{q}_{l_j}$ lies on the nearest plane (or edge) of ${}^{L_k}\hat{\mathbf{p}}_{l_j}$ in local point cloud map. Denoting $\mathbf{u}_j$ as the unit normal vector(or direction vector) of the corresponding plane(or edge), $\mathbf{G}_j = \left[\mathbf{u}_j\right]_\times$ when ${}^{L_k}\mathbf{q}_{l_j} \in \mathbb{F}_e$, and $\mathbf{G}_j = \mathbf{u}_j^T$ when ${}^{L_k}\mathbf{q}_{l_j} \in \mathbb{F}_p$. $[\cdot]_\times$ transfers a vector to its skew-symmetric matrix. The residual defined in Eq. (11) denotes the distance between ${}^{L_k}\hat{\mathbf{p}}_{l_j}$ and its corresponding plane (or edge). We seek for a local pose that minimizes the sum of distances of each point:

$$\Delta\hat{\mathbf{x}}_{L_k} = \underset{\Delta\mathbf{x}_{L_k}}{\arg\min} \sum_{j=1}^{N} \left\| f_j\left(\Delta\mathbf{x}_{L_k}, {}^{L_{k+1}}\mathbf{p}_{l_j}\right)\right\|_2^2 \tag{12}$$

where $\Delta\hat{\mathbf{x}}_{L_k}$ is the estimation of $\Delta\mathbf{x}_{L_k}$ defined in Eq. (3). $\|\cdot\|_2$ denotes the L_2 norm. Eq. (12) can be solved by the Ceres solver [24] using nonlinear Levenberg-Marquardt method.

2) Covariance Estimation:

Inspired by [1], [2], we derive a general 3D ICP-like pose covariance estimation method based on the analysis of the residuals being minimized. We apply the local state covariance for point to edge and point to plane residual. Denoting the error of local state $\Delta\hat{\mathbf{x}}_{L_k}$ as $\widetilde{\mathbf{x}}_k$ ($\Delta\hat{\mathbf{x}}_{L_k} \boxplus \widetilde{\mathbf{x}}_k = \Delta\mathbf{x}_{L_k}$), the covariance of the local state can be calculated as the following equation:

$$\mathrm{cov}\left(\widetilde{\mathbf{x}}_k\right) = \left(\frac{\partial^2 F}{\partial \widetilde{\mathbf{x}}_k^2}\right)^{-1} \frac{\partial^2 F}{\partial \mathbf{z}_{k+1} \partial \widetilde{\mathbf{x}}_k} \mathrm{cov}\left(\mathbf{z}_{k+1}\right) \left(\frac{\partial^2 F}{\partial \mathbf{z}_{k+1} \partial \widetilde{\mathbf{x}}_k}\right)^T \left(\frac{\partial^2 F}{\partial \widetilde{\mathbf{x}}_k^2}\right)^{-1} \tag{13}$$

$$F(\Delta\hat{\mathbf{x}}_{L_k} \boxplus \widetilde{\mathbf{x}}_k, \mathbf{z}_{k+1}) = \sum_{j=1}^{N} \left\| f_j\left(\Delta\hat{\mathbf{x}}_{L_k} \boxplus \widetilde{\mathbf{x}}_k, {}^{L_{k+1}}\mathbf{p}_{l_j}\right)\right\|_2^2 \tag{14}$$

where $\mathbf{z}_{k+1}$ is the stacked vector of LiDAR measurements $\{{}^{L_{k+1}}\mathbf{p}_{l_j}\}$. Supposing the covariance of point measurement is σ^2 and measurements for each point are irrelative, $\mathrm{cov}(\mathbf{z}_{k+1})$ is a diagonal matrix whose diagonal entries are σ^2. For more detailed derivation refer to the supplementary material [1].

E. Pose Measurement Module

In actual implementation of global visual localization, some corner cases, such as the lanes are blocked, sparse, unclear or even repainting, will lead to bad results. Thus, we fuse the relative pose obtained from LO into ESKF for global visual localization.

1) **Pose Measurement Construction:** Construct the global pose measurement by deriving the global updated state at time $k-1$ and relative pose from LO:

$${}^{G}\widetilde{\mathbf{T}}_{B_k} = {}^{G}\overline{\mathbf{T}}_{B_{k-1}} {}^{B}\mathbf{T}_{L} {}^{L_{k-1}}\hat{\mathbf{T}}_{L_k} {}^{B}\mathbf{T}_{L}^{-1} \tag{15}$$

where ${}^{G}\widetilde{\mathbf{T}}_{B_k} \doteq \left({}^{G}\widetilde{\mathbf{R}}_{B_k}, {}^{G}\widetilde{\mathbf{p}}_{B_k}\right)$ denotes the transformation of the constructed global pose from $\mathcal{F}_{B_k}$ to $\mathcal{F}_G$. ${}^{G}\overline{\mathbf{T}}_{B_{k-1}}$ denotes transformation of the updated state $\overline{\mathbf{x}}_{k-1}$ obtained by ESKF at time $k-1$ in Sec. III-F. ${}^{L_{k-1}}\hat{\mathbf{T}}_{L_k}$ is the transformation of the local state $\Delta\hat{\mathbf{x}}_{L_{k-1}}$ estimated in Sec. III-D.1.

The residuals defined the differences in translation and attitude between the estimated pose and pose measurement:

$$\mathbf{z}_p = {}^{G}\hat{\mathbf{p}}_{B_k} - {}^{G}\widetilde{\mathbf{p}}_{B_k} \tag{16a}$$

$$\mathbf{z}_r = \log({}^{G}\widetilde{\mathbf{R}}_{B_k}^{-1} {}^{G}\hat{\mathbf{R}}_{B_k})^\vee \tag{16b}$$

where $\log(\cdot)^\vee$ is the logarithmic map from $\mathrm{SO}(3) \to \mathfrak{so}(3)$. ${}^{G}\hat{\mathbf{R}}_{B_k}$ is the estimated rotation from $\mathcal{F}_{B_k}$ to $\mathcal{F}_G$ at time k.

[1]https://github.com/llpan91/Supplementary-Material-to-VILC

2) **Relative Pose Observation Model:** Ideally, the translation and rotation in Eq. (15) represent the ground-truth position and attitude of IMU in $\mathscr{F}_G$. However, ignore the noise from calibrated extrinsic parameters, the errors of ${}^G\widetilde{\mathbf{p}}_{C_k}$ and ${}^G\widetilde{\mathbf{R}}_{C_k}$ originate from the updated state and local state. Considering the errors of the updated state and local state leads to the ground-truth position ${}^G\mathbf{p}^{\text{gt}}_{B_k}$ and attitude ${}^G\mathbf{R}^{\text{gt}}_{B_k}$:

$$\begin{aligned}{}^G\mathbf{p}^{\text{gt}}_{B_k} =& \Psi\left({}^G\widetilde{\mathbf{p}}_{B_k}, \boldsymbol{\nu}_k\right) \\ =& {}^G\overline{\mathbf{R}}_{B_{k-1}} \exp(\delta\boldsymbol{\theta}_{k-1})({}^B\mathbf{R}_L({}^{L_{k-1}}\hat{\mathbf{R}}_{L_k} \exp(\delta\widetilde{\boldsymbol{\theta}}_{k-1})^L\mathbf{p}_B \\ &+ {}^{L_{k-1}}\hat{\mathbf{p}}_{L_k} + \delta\widetilde{\mathbf{p}}_{k-1}) + {}^B\mathbf{p}_L) + {}^G\overline{\mathbf{p}}_{B_{k-1}} + \delta\mathbf{p}_{k-1}\end{aligned} \tag{17}$$

$$\begin{aligned}{}^G\mathbf{R}^{\text{gt}}_{B_k} =& \Phi({}^G\widetilde{\mathbf{R}}_{B_k}, \boldsymbol{\nu}_k) \\ =& {}^G\overline{\mathbf{R}}_{B_{k-1}} \exp(\delta\boldsymbol{\theta}_{k-1}){}^B\mathbf{R}_L{}^{L_{k-1}}\hat{\mathbf{R}}_{L_k} \exp(\delta\widetilde{\boldsymbol{\theta}}_{k-1}){}^B\mathbf{R}_L^{-1}\end{aligned} \tag{18}$$

where $\delta\boldsymbol{\theta}_{k-1}$, $\delta\mathbf{p}_{k-1}$, $\delta\widetilde{\boldsymbol{\theta}}_{k-1}$, $\delta\widetilde{\mathbf{p}}_{k-1}$ are errors between ${}^G\overline{\mathbf{R}}_{B_{k-1}}$, ${}^G\overline{\mathbf{p}}_{B_{k-1}}$, ${}^{L_{k-1}}\hat{\mathbf{R}}_{L_k}$, ${}^{L_{k-1}}\hat{\mathbf{p}}_{L_k}$ and the corresponding ground-truth state, respectively. $\exp(\cdot)$ is the exponential map from $\mathbb{R}^3 \rightarrow$ SO(3). $\boldsymbol{\nu}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}_k})$, which is composed of the errors above, can be described as:

$$\boldsymbol{\nu}_k = \begin{bmatrix}[\delta\mathbf{x}_{k-1}]^T_{3,1} & \widetilde{\mathbf{x}}^T_{k-1}\end{bmatrix}^T = \begin{bmatrix}\delta\boldsymbol{\theta}^T_{k-1} & \delta\mathbf{p}^T_{k-1} & \delta\widetilde{\boldsymbol{\theta}}^T_{k-1} & \delta\widetilde{\mathbf{p}}^T_{k-1}\end{bmatrix}^T \tag{19}$$

where $[\cdot]_{3,1}$ denotes the third and first components of a vector i.e. $[\delta\mathbf{x}_{k-1}]_{3,1} = \begin{bmatrix}\delta\boldsymbol{\theta}^T_{k-1} & \delta\mathbf{p}^T_{k-1}\end{bmatrix}^T$. $\widetilde{\mathbf{x}}_{k-1} = \begin{bmatrix}\delta\widetilde{\boldsymbol{\theta}}^T_{k-1} & \delta\widetilde{\mathbf{p}}^T_{k-1}\end{bmatrix}^T$ is the error of local state. Assuming $[\delta\mathbf{x}_{k-1}]_{3,1}$ and $\widetilde{\mathbf{x}}_{k-1}$ are irrelevant leads to the block diagnal matrix $\boldsymbol{\Sigma}_{\boldsymbol{\nu}_k}$:

$$\boldsymbol{\Sigma}_{\boldsymbol{\nu}_k} = \text{diag}\left(\text{cov}\left([\delta\mathbf{x}_{k-1}]_{3,1}\right), \text{cov}\left(\widetilde{\mathbf{x}}_{k-1}\right)\right) \tag{20}$$

where $\text{cov}\left([\delta\mathbf{x}_{k-1}]_{3,1}\right)$ can be obtained from the covariance matrix $\overline{\mathbf{P}}_{k-1}$ of the updated state $\overline{\mathbf{x}}_{k-1}$ by ESKF. $\text{cov}\left(\widetilde{\mathbf{x}}_{k-1}\right)$ is the relative pose covariance introduced in Sec. III-D.2.

The ground-truth pose measurement with the ground-truth state $\mathbf{x}_k$ should lead to zero in Eq. (16). For example, the IMU position in the ground-truth state $\mathbf{x}_k$ should be the same as ${}^G\mathbf{p}^{\text{gt}}_{B_k}$ in which errors of ${}^G\widetilde{\mathbf{p}}_{B_k}$ are considered:

$$\begin{aligned}\mathbf{0} = \mathbf{h}_p(\mathbf{x}_k, \boldsymbol{\nu}_k) &= {}^G\mathbf{p}_{B_k} - \Psi\left({}^G\widetilde{\mathbf{p}}_{B_k}, \boldsymbol{\nu}_k\right) \\ &\simeq \mathbf{h}_p(\hat{\mathbf{x}}_k, \mathbf{0}) + \mathbf{H}_p\delta\mathbf{x}_k + \mathbf{J}_p\boldsymbol{\nu}_k = \mathbf{z}_p + \mathbf{H}_p\delta\mathbf{x}_k + \mathbf{J}_p\boldsymbol{\nu}_k\end{aligned} \tag{21}$$

where $\mathbf{H}_p$ is the Jacobian matrix of $\mathbf{h}_p(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \boldsymbol{\nu}_k)$ with respect to $\delta\mathbf{x}_k$. $\mathbf{J}_p$ is the Jaccobian matrix of $\mathbf{h}_p(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \boldsymbol{\nu}_k)$ with respect to $\boldsymbol{\nu}_k$.

Similar to Eq. (21), the ground-truth attitude constructed in Eq. (18) should be the same as the attitude of IMU with the ground-truth state $\mathbf{x}_k$. The pose measurement model for attitude can be described as:

$$\begin{aligned}\mathbf{0} = \mathbf{h}_r(\mathbf{x}_k, \boldsymbol{\nu}_k) &= \log\left(\Phi({}^G\widetilde{\mathbf{R}}_{B_k}, \boldsymbol{\nu}_k)^{-1} \cdot {}^G\mathbf{R}_{B_k}\right)^\vee \\ &\simeq \mathbf{h}_r(\hat{\mathbf{x}}_k, \mathbf{0}) + \mathbf{H}_r\delta\mathbf{x}_k + \mathbf{J}_r\boldsymbol{\nu}_k = \mathbf{z}_r + \mathbf{H}_r\delta\mathbf{x}_k + \mathbf{J}_r\boldsymbol{\nu}_k\end{aligned} \tag{22}$$

where $\mathbf{H}_r$ and $\mathbf{J}_r$ are the Jacobian matrices of $\mathbf{h}_r(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \boldsymbol{\nu}_k)$ with respect to $\delta\mathbf{x}_k$ and $\boldsymbol{\nu}_k$ respectively.

F. State Estimation

We have discussed four types of the measurement model. The visual measurement models for dashed lanes in Eq. (7) and solid lanes in Eq. (9) can provide constraints to perform global localization. The pose measurement models for the position in Eq. (21) and attitude in Eq. (22) can provide constraints between two frames to enhance the accuracy and robustness of global localization.

Because the camera and LiDAR are synchronized by hardware, we consider the point cloud frame and image frame can be obtained simultaneously. Once we receive valid measurements, we can estimate the updated state $\overline{\mathbf{x}}_k$. Writing the four types of measurement model in compact form $\mathbf{h}(\mathbf{x}_k, \mathbf{n}_k)$, which is the stacked vector of $\{\mathbf{h}_{d_i}(\cdot)\}$, $\{\mathbf{h}_{s_j}(\cdot)\}$, $\mathbf{h}_p(\cdot)$ and $\mathbf{h}_r(\cdot)$. We can estimate the error state $\delta\mathbf{x}_k$ by the following equations:

$$\begin{aligned}\mathbf{K} &= \hat{\mathbf{P}}_k\mathbf{H}_k^T(\mathbf{H}_k\hat{\mathbf{P}}_k\mathbf{H}_k^T + \mathbf{J}_k\Sigma_{\mathbf{n}_k}\mathbf{J}_k^T) \\ \delta\mathbf{x}_k &= -\mathbf{K}\mathbf{h}(\hat{\mathbf{x}}_k, \mathbf{0}) \\ \overline{\mathbf{P}}_k &= (\mathbf{I} - \mathbf{K}\mathbf{H})\hat{\mathbf{P}}_k\end{aligned} \tag{23}$$

where $\Sigma_{\mathbf{n}_k}$ is the covariance of $\mathbf{n}_k$ which is the stacked vector of $\{\mathbf{n}_{d_i}\}$, $\{\mathbf{n}_{s_j}\}$ and $\boldsymbol{\nu}_k$. $\mathbf{h}(\hat{\mathbf{x}}_k, \mathbf{0})$ is the stacked vector of $\{\mathbf{z}_{d_i}\}$, $\{\mathbf{z}_{s_j}\}$, $\mathbf{z}_p$ and $\mathbf{z}_r$. $\mathbf{J}_k$, composed of $\mathbf{J}_p$ and $\mathbf{J}_r$, is the Jaccobian matrix of $\mathbf{h}(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \mathbf{n}_k)$ w.r.t $\mathbf{n}_k$. $\mathbf{H}_k$, composed of $\{\mathbf{H}_{d_i}\}$, $\{\mathbf{H}_{s_j}\}$, $\mathbf{H}_p$ and $\mathbf{H}_r$, is the Jaccobian matrix of $\mathbf{h}(\hat{\mathbf{x}}_k \boxplus \delta\mathbf{x}_k, \mathbf{n}_k)$ w.r.t $\delta\mathbf{x}_k$.

As the ESKF tradition, we can obtain the updated state $\overline{\mathbf{x}}_k$ after the error state $\delta\mathbf{x}_k$ is resolved and set $\hat{\mathbf{x}}_k = \overline{\mathbf{x}}_k$ until the subsequent IMU measurement is received.

IV. EXPERIMENTS

We construct two experiments to evaluate the our proposed algorithm's performance in this section. We collect experimental data from three routes, including highway, sea-crossing bridge and tunnel. There are many challenging cases in the routes, such as lane missing, lane occlusion, LiDAR degradation, etc. During testing and evaluation, a LiDAR and camera are used to collect image and point clouds of the environment, respectively. IMU collects vehicle motion information, including acceleration and angular velocity with high frequency. Besides, the results of Novatel (integrating RTK with GNSS) are fused with LiDAR, wheel encoder, and IMU in an optimized framework to generate ground truth.

A. Localization Performance

The evaluation metrics of localization accuracy usually include lateral error, longitudinal error, and yaw error. We further evaluate the accuracy of pitch and roll. In addition to positioning accuracy, the robustness of localization is even more critical for safe autonomous driving tasks. We apply the Three Sigma Limits to describe the robustness, which indicates that 93.3% errors are within the interval of 3σ around the mean. The raw error distribution is shown in Fig. 5, and statistics of absolute error are shown in Table II.

In our test scenarios, lateral errors are within 12.5 cm in most cases, and the mean of **absolute lateral error** is less

TABLE II

STATISTICS OF ABSOLUTE POSITION AND ATTITUDE ERROR ON REAL-WORLD TEST SEQUENCES

Scenes	Length	Lateral Error(m)		Longitudinal Error(m)		Pitch Error(deg)		Roll Error(deg)		Yaw Error(deg)	
		mean	3σ	mean	3σ	mean	3σ	mean	3σ	mean	3σ
Highway	30.2km	0.048	0.086	0.162	0.375	0.088	0.213	0.329	1.28	0.070	0.240
Sea-crossing Bridge	26.5km	0.045	0.117	0.149	0.537	0.089	0.211	0.280	1.146	0.198	0.217
Tunnel	0.96km	0.059	0.117	0.158	0.483	0.102	0.252	0.618	2.24	0.067	0.257

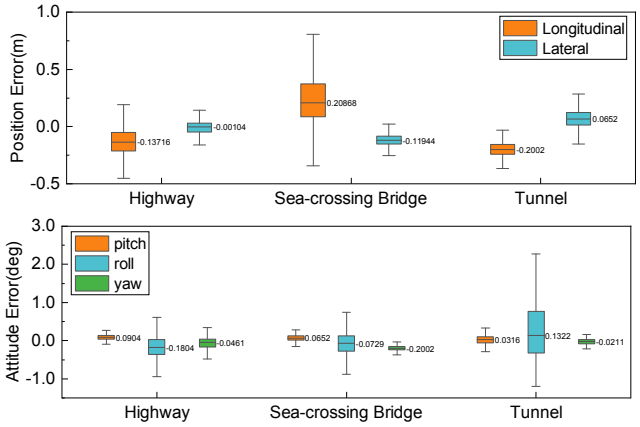
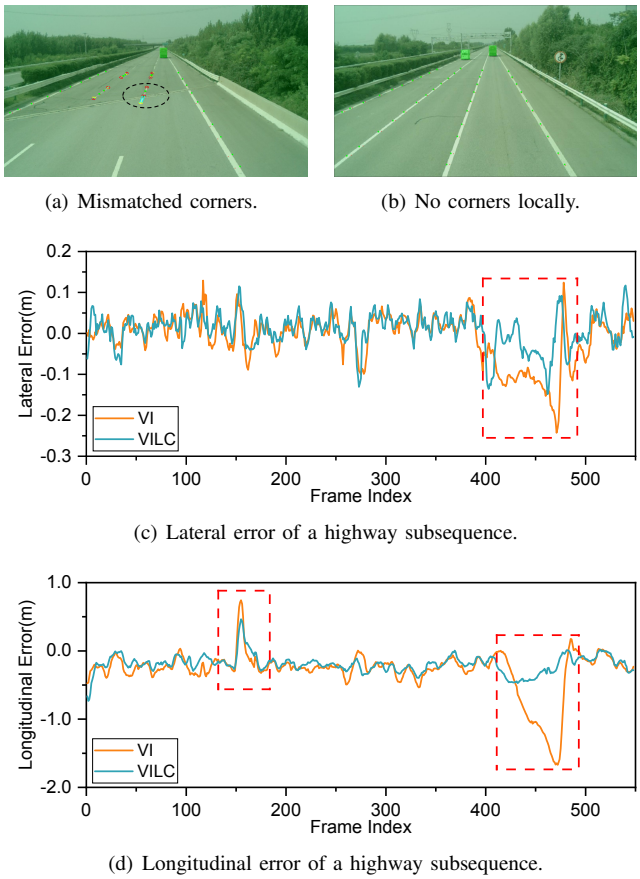


Fig. 5. Box plot of raw position and attitude errors of 3 routes using the proposed method **VILC.**



(a) Mismatched corners. (b) No corners locally.

(c) Lateral error of a highway subsequence.

(d) Longitudinal error of a highway subsequence.

Fig. 6. Position errors of a highway subsequence with mismatched corners and solid lanes only.

than **5.9 cm**. The mean of **absolute longitudinal error** is less than **15.8 cm**. The mean **absolute attitude error** of all scenes are within **0.102**°, **0.618**° and **0.198**° for pitch, roll and yaw respectively. The roll error and uncertainty are larger than pitch and yaw. The results are reasonable because the observations from the lane are all planar, which under-constraint to roll.

B. Ablation Study

We perform an ablation study to evaluate the contributions of the visual-inertial localization and LO with pose covariance estimation in our proposed method. Different settings are defined as below. **VI** uses the visual-inertial localization method with semantic map. **VIL** fuses the LO into the visual-inertial localization method, but without estimating the covariance of relative poses in real-time. **VILC** is our proposed method, which fuses the LO with online covariance into the visual-inertial localization system. The results are shown in Table III. It can be seen that **VILC** achieves the best localization performance and robustness.

In more detail, we demonstrate the robustness of our proposed method **VILC** compared with **VI** in real-world scenarios. Fig. 6a shows the case that dashed lane repainting cause keypoints mismatching. Fig. 6b shows the case that dash corners are missing due to extensive abrasion and cause longitudinal under-constrained. Fig. 6c and Fig. 6d show the raw lateral error and longitudinal error of the above case scenario, respectively. These two cases occurred around frames 150 and 400. As Fig. 6c shown, there exists a lateral drift when only solid lanes are observed (during frame 400 to 500), and this phenomenon also appears in [11], [13].

Our proposed method has a smaller drift because the LO can provide extra relative constraints. As Fig. 6d shown, there exits a sudden peak at the frame 150 due to corners mismatching. Our proposed method can handle these cases due to fused drift-aware LO, which can reduces the longitudinal error when the mismatching occurred. During frame 400 to 500, for **VI** the longitudinal error accumulates until the corners of dashed lines are observed at frame 470, while our proposed method maintains an excellent longitudinal localization performance.

C. Runtime

We test the runtime of our proposed method on a computing platform (a desktop with Core i7-7700K CPU@4.2GHz and 16GB RAM). Table IV lists the runtime of each module in our proposed method, where the abbreviations VMM,

TABLE III

STATISTICS OF ABSOLUTE ERROR FOR THE ABLATION STUDY

Scenes	Setting	Lat. Error(m)		Long. Error(m)		Yaw Error(deg)	
		mean	3σ	mean	3σ	mean	3σ
Highway	VI	0.050	0.141	0.201	0.729	0.076	0.269
	VIL	0.050	0.135	0.190	0.483	0.076	0.252
	VILC	**0.048**	**0.086**	**0.162**	**0.375**	**0.070**	**0.240**
Bridge	VI	0.047	0.129	0.152	0.546	0.200	0.223
	VIL	0.046	0.125	0.163	0.764	**0.198**	0.228
	VILC	**0.045**	**0.117**	**0.149**	**0.537**	**0.198**	**0.217**
Tunnel	VI	0.062	0.138	0.212	0.558	0.072	0.261
	VIL	0.063	0.137	0.186	0.535	0.069	0.258
	VILC	**0.059**	**0.117**	**0.158**	**0.483**	**0.067**	**0.257**

TABLE IV

RUNTIME OF EACH MODULE

Total	VMM	LOM	PMM	ESKF
41.22ms	19.75ms	20.5ms	0.14ms	0.83ms

LOM, and PMM stand for the Visual Measurement Module, LO Module, and Pose Measurement Module, respectively. The results show that our proposed method can perform a real-time localization.

V. CONCLUSIONS

This paper developed an ESKF-based visual localization system that exploits fused global semantic observation constraints and relative pose constraints to estimate vehicle pose. Our method fully explores the complementarity of camera and LiDAR, global and relative constraints. Also, our visual measurements model contains keypoint observation constraints and distance transform constraints to guarantee the system's accuracy and robustness. Future work is to explore more stable semantic elements (such as traffic signs, poles, etc.) to further improve the system.

REFERENCES

[1] A. Censi, "An accurate closed-form estimate of icp's covariance," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 3167–3172, 2007.

[2] S. M. Prakhya, L. Bingbing, Y. Rui, and W. Lin, "A closed-form estimate of 3d icp covariance," in *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 526–529, 2015.

[3] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[4] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Pl-slam: Real-time monocular visual slam with points and lines," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 4503–4508, IEEE, 2017.

[5] L. Pan, J. Cheng, W. Feng, and X. Ji, "A robust rgb-d image-based slam system," in *International Conference on Computer Vision Systems*, pp. 120–130, Springer, 2017.

[6] W. Lee, K. Eckenhoff, P. Geneva, and G. Huang, "Intermittent gps-aided vio: Online initialization and calibration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5724–5731, 2020.

[7] R. Mascaro, L. Teixeira, T. Hinzmann, R. Siegwart, and M. Chli, "Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1421–1428, 2018.

[8] J. Liu, W. Gao, and Z. Hu, "Optimization-based visual-inertial slam tightly coupled with raw gnss measurements," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11612–11618, 2021.

[9] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. 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.

[10] S. Se, D. Lowe, and J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 364–375, 2005.

[11] M. Schreiber, C. Knöppel, and U. Franke, "Laneloc: Lane marking based localization using highly accurate maps," in *2013 IEEE Intelligent Vehicles Symposium (IV)*, pp. 449–454, 2013.

[12] T. Wu and A. Ranganathan, "Vehicle localization using road markings," in *2013 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1185–1190, 2013.

[13] C. Zhang, H. Liu, H. Li, K. Guo, K. Yang, R. Cai, and Z. Li, "Dt-loc: Monocular visual localization on hd vector map using distance transforms of 2d semantic detections," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1531–1538, 2021.

[14] W.-C. Ma, I. Tartavull, I. A. Bârsan, S. Wang, M. Bai, G. Mattyus, N. Homayounfar, S. K. Lakshmikanth, 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)*, pp. 5304–5311, 2019.

[15] 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)*, pp. 11255–11261, 2021.

[16] W. Xu and F. Zhang, "Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter," *IEEE Robotics and Automation Letters*, 2021.

[17] C. Hertzberg, R. Wagner, U. Frese, and L. Schröder, "Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds," *Information Fusion*, vol. 14, no. 1, pp. 57–77, 2013.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[19] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.

[20] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[21] G. Borgefors, "Distance transformations in digital images," *Computer vision, graphics, and image processing*, vol. 34, no. 3, pp. 344–371, 1986.

[22] J. Ruyi, K. Reinhard, V. Tobi, and W. Shigang, "Lane detection and tracking using a new lane model and distance transform," *Machine vision and applications*, vol. 22, no. 4, pp. 721–737, 2011.

[23] J. Zhang and S. Singh, "Loam : Lidar odometry and mapping in real-time," pp. 109–111, 01 2014.

[24] S. Agarwal, K. Mierle, and T. C. S. Team, "Ceres Solver," 3 2022.