Accurate and Robust Visual Localization System in Large-Scale Appearance-Changing Environments

Yang Yu, Peng Yun, Bohuan Xue, *Graduate Student Member, IEEE*, Jianhao Jiao, Rui Fan, *Member, IEEE*, and Ming Liu, *Senior Member, IEEE*

***Abstract*—Localization in large-scale environments with robust performance is a persistent challenge for mobile robots. This article proposes a novel system to achieve accurate and robust visual localization performance in large-scale environments with appearance-changing surroundings. Our system starts from a stage of extracting stable visual features with an object segmentation network. After measurement postprocessing and extrinsic precalibration, we propose a graph-based optimization module to estimate the optimal pose as well as extrinsics. We construct optimization constraints with refined wheel odometry, feature matching between images, and correspondences between images and the prebuild map. We evaluate our segmentation module on our proposed datasets and test our localization module with seven sequences (9.8 km total length) in real port scenes with different working conditions from day to night and sunny to rainy. Experiment results demonstrate the decimeter-level accuracy and robust performance of our approach in various challenging scenarios, showing competitive performance compared with state-of-the-art LiDAR-based localization methods.**

***Index Terms*—Graph-based optimization, neural network, online calibration, visual localization.**

Manuscript received 20 January 2022; accepted 11 May 2022. Date of publication 3 June 2022; date of current version 14 December 2022. Recommended by Technical Editor S. G. Loizou and Senior Editor H. Qiao. This work was supported by Zhongshan Science and Technology Bureau Fund under Grant 2020AG002, in part by the Foshan-HKUST under Grant FSUST20-SHCIRI06C, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A0505090008, awarded to Prof. Ming Liu. *(Corresponding author: Ming Liu.)*

Yang Yu, Peng Yun, Bohuan Xue, and Jianhao Jiao are with the Department of Electrical and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: yyubj@connect.ust.hk; pyun@connect.ust.hk; bxueaa@ust.hk; jjiao@ust.hk).

Rui Fan is with the Department of Control Science and Engineering, Tongji University, Shanghai 201804, China (e-mail: ranger_fan@outlook.com).

Ming Liu is with the Department of Electrical and Computer Engineering, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China, with the The Hong Kong University of Science and Technology, Hong Kong, Hong Kong, and also with the HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China (e-mail: eelium@ust.hk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMECH.2022.3177237.

Digital Object Identifier 10.1109/TMECH.2022.3177237

I. Introduction

A. Motivation

In Recent years, demands of deploying unmanned ground vehicles (UGVs) in limited areas, such as unmanned port vehicles (UPVs) in ports, are growing. Accurate localization (error $\leq$ 10 cm) of UPVs for containers loading is the most fundamental ability. The GPS, however, is commonly interfered by cranes' mechanical structure. With the development of simultaneous localization and mapping (SLAM) in autonomous driving, visual localization approaches show their potential for such challenging tasks. Existing localization approaches with a monocular camera have been demonstrated with great properties (e.g., low cost and small size) on various platforms such as drones [1], [2] and UGVs [3]. However, they still suffer from shortcomings such as appearance changes in outdoor environments [4]. For instance, illumination and seasonal appearance change extremely affect image perception tasks [5]. In addition, existing visual localization algorithms [3], [6] usually assume a working environment with static surroundings (e.g., trees, buildings) for corresponding matching between descriptors and a 3-D map. This characteristic usually induces significant outliers to data associations in real-world environments, such as port scenes, which only have frequently moving container stacks, trucks, and gantry cranes. Last but not least, most visual localization approaches have meter-level drifts after several kilometers of movement, which is not acceptable for autonomous driving applications.

Recently, we have noticed the outstanding performance of convolutional neural networks (CNNs) in challenging perceptual conditions [7]–[10], and the trend of detecting predefined semantic visual markers [11] instead of the traditional point features in large-scale localization tasks. Compared with the traditional monocular visual-based localization approach, we consider that a multicamera system with neural networks and semantic auxiliaries enhances system robustness significantly.

B. Challenges

Although visual-based approaches have shown remarkable performance for accurate localization, there are still several challenges in the large-scale outdoor environments:

1) Localization Accuracy: To ensure the accuracy of the visual simultaneous localization and mapping (SLAM) system, pose estimation, and mapping algorithms are compounded together for consistency. Accumulated drifts of pose estimation and map construction are increasing with the movement of robots, and loop closure eliminates drifts by recognizing already visited areas compared with the 3-D map. For better system accuracy, the prebuild point cloud map is adopted instead of constructing in real-time [3]. However, this approach is only available in scenes with static references such as buildings and trees for prebuild map consistency. In port scenes, changing appearance (container stacks and gantry cranes) affects the consistency of the maintained map, leading to mismatching and failure in the loop closure. In addition, computational resources also limit the size of the constructing map in large-scale environments.

2) Extrinsic Calibration: As emphasized in our previous works [12], the slight extrinsic perturbation is detrimental to the system performance. Heavy containers affect UPVs' mechanical structure dramatically, leading to variations of sensors' extrinsic parameters. Offline hand-eye calibration [13] is not available for real-time systems. Online calibration problems have been studied in [2] by simultaneous estimation of extrinsics and ego-motion in a tightly coupled optimization module. However, this process easily fails for local optimized results if systems do not have a good initial value of extrinsics. In order to manage the centimeter-level localization error, it is essential to model the extrinsic perturbation online.

3) Appearance Changing: Visual-based systems suffer from appearance-changing problems a lot. Illumination, seasonal appearance, and weather conditions variation lead to unstable descriptors extraction. Learning-based approaches can handle challenging perceptual conditions, but the large computational consumption and additional acceleration computational units limit their application in industry scenes. In order to meet this tradeoff, we need to design an efficient network structure for real-time feature extraction.

C. Contributions

These challenges motive us to propose a novel visual localization approach in large-scale appearance-changing environments, such as port scenes. We consider a lightweight and globally consistent map instead of the popular point cloud map or building the map in real-time. The map only comprises 2-D positions of predefined markers and lanes on the ground, which are more common and consistent than natural structural objects in ports. To enhance the robustness of our system under various appearance conditions, we propose a lane and diamond-shaped road marker segmentation network (LDS-Net). In addition, we propose a complete camera-vehicle calibration module to solve the problem of extrinsic variations affected by dynamic loads. We first introduce the precalibration step with the vehicle structure model and epipolar constraints of images before and after container loading. With these inputs, an accurate and robust localization module is then proposed with a graph-based optimization module.

Accordingly, we identify our contributions as follows:

1) An efficient network structure for lane and marker segmentation. We extend spatial convolution neural network (SCNN) [9], and Mask-RCNN [8] with a shared backbone for fast and efficient performance.
2) A complete self-calibration module to eliminate the effects of UPVs' loading variations. We utilize a vehicle-structure-based model as an extrinsics initial value and optimize it with the direct method of epipolar constraints.
3) A novel visual localization system for large-scale environments with appearance-changing surroundings. We construct a lightweight prebuild map and a graph-based optimization approach that utilizes the refined wheel odometry preintegration, feature matching of detected masks, and PnP-based optimization for a robust and accurate localization performance.

We validate the proposed system in various working conditions, showing the decimeter-level accuracy and long-duration consistency, which is competitive to state-of-the-art LiDAR-based methods. To the best of our knowledge, our proposed system is the first visual-based localization solution for UPVs applied in real port operations. To benefit the research community, we release an open-source port scene image dataset,[1] including semantic information such as lane and road marker information in various working conditions.

D. Organization

In Section V, we propose the architecture of the LDS-Net. RCNN-based [8] object segmentation generates the road marker mask, and the SCNN-based [9] lane segmentation provides the lane information, in Sections V-B and V-C, respectively. We present the measurement preprocessing in Section VI. We also consider the extrinsic revision by loading variation in Section VII. Graph-based optimization is introduced in VIII. And finally, Section IX concludes this article.

II. Related Work

A. Large-Scale Visual Localization

Image-retrieval methods regress camera poses by searching out the most high-ranking stored image and the camera pose from the prior database [14], [15]. The efficiency and accuracy highly depend on 2-D descriptor selections and database sampling strategies. To reduce false-positive matches and influence by appearance changes, Milford and Wyeth [16] matched a sequence of images rather than an individual image. With the development of CNN, learning-based visual descriptors strengthen the place recognition performance by dense pixel-wise features or semantic understanding [17]–[19]. Daniel [20] adopted sparse learned features, including key points and descriptors to replace counterparts.

Compared with image-retrieval approaches, structure-based approaches have been shown to have a great improvement in localization accuracy. 2-D descriptors are extracted to match

[1] [Online]. Available: https://sites.google.com/view/dsl-dataset

with the 3-D dense, or sparse map by RANSAC and PnP solver [21]. Huang *et al.* [22] proposed a Gaussian Mixture-based representation to extract compact structural features from the map. Such methods [23], [24] show great performance only in environments with static and structural reference such as buildings or trees. With map size growing, this matching process becomes inefficient and computational resource-consuming.

To realize the visual localization in appearance changing scenes, applying predefined road markers on the ground, which are consistent and friendly to the construction, is a favorable option for robust feature extraction. Therefore, we propose a visual localization system with road markers segmentation and a graph-based optimization modules.

B. Object Detection and Segmentation

There is plentiful existing research on object detection and segmentation based on deep neural networks [10], [25]–[27]. They adopt CNN architectures to extract meaningful features and subsequently exploit these features for classification and regression tasks [28]. In recent years, with the inspiration of feature pyramid, FPN is proposed to extract multiscale features from images and improve the scale invariance [29].

The region-proposal framework is a popular pipeline for object detection. It first generates proposals with unsupervised methods [30] or a region proposal network (RPN) [25], and then classifies and refines the proposals with SVMs or CNNs. Compared with object detection, instance segmentation networks estimate additional masks for detected bounding boxes [8], [31]. In our method, we adopt the Mask R-CNN architecture [8] for diamond road marker segmentation.

Long *et al.* first proposed deep learning networks for semantic segmentation with a full convolution network (FCN), which includes only convolutional layers [32]. It allows the network to take an image of arbitrary size and estimate pixel-wise classes of the same shape. Huval *et al.* [33] first attempted adopting deep learning in lane detection. Pan *et al.* [9] proposed SCNN for lane segmentation under the context of autonomous driving and innovated spatial convolution layers to embed the spatial information of pixels into feature space. In our proposed LDS-Net, we solve a multitask learning problem and jointly achieve road marker segmentation and lane segmentation. An FPN is adopted to extract shared multiscale features for these two tasks.

III. Problem Statement

We denote the pose of the UPV as $\mathbf{x}_k$ in the kth frame. The set of observations $\boldsymbol{z}_k = (\boldsymbol{z}_{k1}, \ldots \boldsymbol{z}_{kn})$ contains n random variables for multiple sensors inputs. $\mathbf{x}_k$ can be derived as

$$\mathbf{x}_k^* = \arg\min_{\mathbf{x_k}} P(\boldsymbol{z_k}|\mathbf{x}_k)P(\mathbf{x}_k) = \arg\min_{\mathbf{x_k}} \prod_{i=1}^{n} P(\boldsymbol{z_{ki}}|\mathbf{x}_k)P(\mathbf{x}_k). \tag{1}$$

This equation can be treated as a maximum likelihood estimation (MLE) to compute the most likely pose that maximized the pdf with a least-squares problem

$$\mathbf{x}_k^* = \arg\min_{\mathbf{x_k}} \frac{1}{2} \sum_{i=1}^{n} \parallel \boldsymbol{N}_i(\mathbf{x}_k) - \boldsymbol{z_{ki}} \parallel_{\Sigma_i}^2 \tag{2}$$

where $\parallel \boldsymbol{a} \parallel_{\Sigma}^2 = \boldsymbol{a}^\top \Sigma^{-1} \boldsymbol{a}$ and Σ denotes the covariance matrix of residuals. We extend the MLE to our visual localization system for the pose estimation and extrinsics in (6) and (7).

TABLE I
NOMENCLATURE

Notation	Explanation
$\mathbf{F}$	Frame, where $\mathbf{F}_W$, $\mathbf{F}_b$, and $\mathbf{F}_c$ represent the world frame, the UPV baselink, and the camera frame;
$\mathbf{T}$	Transform matrix in SE(3), where $\mathbf{T}_b^a$ transforms a point x_b in $\mathbf{F}_b$ into the $\mathbf{F}_a$;
$\mathbf{R}$	Rotation matrix in $SO(3)$, where $\mathbf{R}_b^a$ represents the rotation from $\mathbf{F}_b$ to $\mathbf{F}_a$;
$\mathbf{q}$	Quaternion under Hamilton notation, with $\mathbf{q}_b^a$ corresponding to $\mathbf{R}_b^a$;
$\mathbf{t}$	Translation vector in $\mathbb{R}^3$, where $\mathbf{t}_b^a$ represents the translation, from $\mathbf{F}_b$ to $\mathbf{F}_a$ defined in $\mathbf{F}_a$;
$\mathbf{P}$	A road marker edge intersection $[x, y, z]^\top$ in $\mathbf{F}_W$;
$\mathbf{p}$	A detected key feature point $[u, v, 1]^\top$ in image plane;
$\mathcal{L}$	The loss function of the proposed network;
ϵ	The cross-entropy function of the proposed network;
$\mathbf{L}$	The final localization result.

IV. Overview

Our proposed visual localization pipeline is shown in Fig. 1. The network provides segmentation results of road markers and lanes for extracting edge intersections and the heading angle. The wheel odometry and heading angle are adopted to provide a higher frequency but easy-to-drift refined odometry. An optimization module utilizes a graph-based approach for pose and extrinsic optimization for reducing accumulated drifts. We also analyze the influence of extrinsic perturbation by formulating it as a polynomial fitting model as prior and epipolar constraints from homograph transformation.

The nomenclature is defined in Table I. $\mathbf{F}_b$ is the UPV baselink with no loadings. We assume that the proposed system is for UPVs, so we only consider the 2-D translation estimation in xy plane and yaw angle in rotation estimation. Due to the large size of the UGV (15 $\times$ 3 $\times$ 1.7 m), we utilize two cameras, which are installed on the front and rear part, respectively, of the vehicle for better heading angle estimation. Intrinsic parameters of cameras are assumed to be known with pitch angle revising to the ground for better road marker observation. We store all diamond-shaped road marker corners' and lanes' 3-D positions measured by the electronic total station as a lightweight prebuild map. We assume that the initial pose of UPV is assigned by the GPS with road marker observation. Initial extrinsic is calculated by 2D–3D matching of road marker mask and prebuild map. We also assume that cameras are rigidly connected with the UPV.

V. Road Marker and Lane Segmentation

Our proposed network takes two raw images, each from the front and rear monoculars, as inputs. Outputs are masks

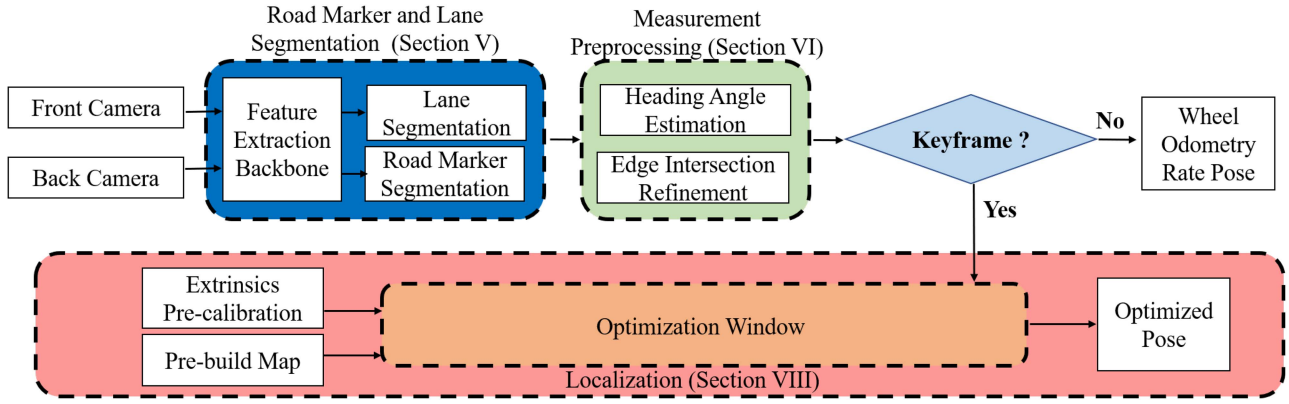


Fig. 1. Pipeline of the proposed system. The system starts with feature extraction and two segmentation tasks in Section V. We refine edge intersections and heading angle information in Section VI. The graph-based optimization consists of refined wheel odometry, keyframes feature matching, and PnP corresponding (see Section VIII). Keyframes are images with detected road markers.

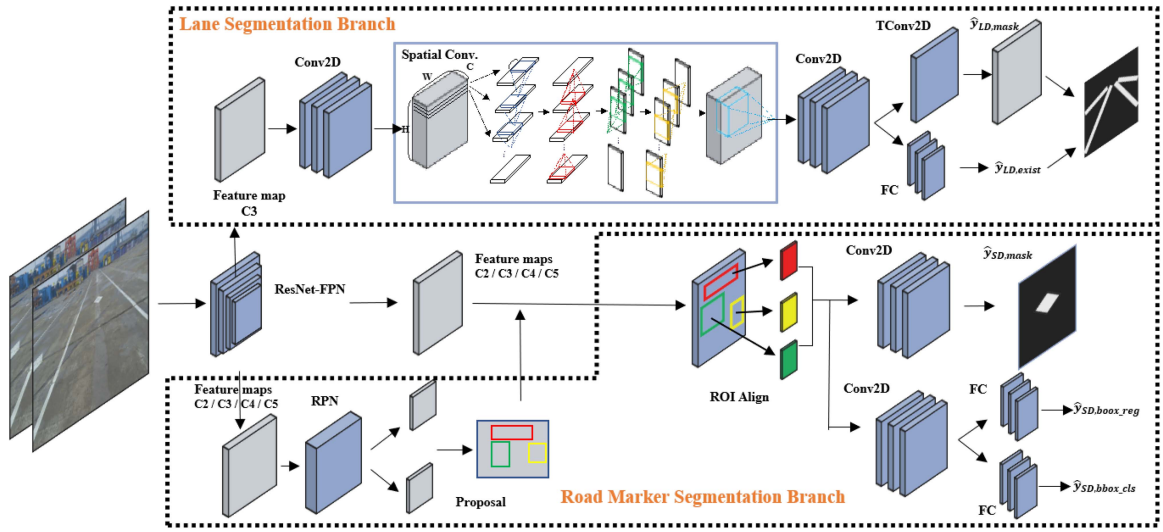


Fig. 2. LDS-Net architecture. It contains three modules: ResNet-FPN for multiscale feature extraction, a lane segmentation branch, and a road marker segmentation branch. Blue blocks denote network layers, while gray blocks denote features in embedding spaces (best viewed in color).

of road markers and lanes. Fig. 2 shows the overview of network architecture with three modules: feature extractor with feature pyramid network (FPN) backbone, diamond-shaped road marker segmentation, and lane segmentation.

A. FPN-Based Feature Extractor

Diamond-shaped road markers span the whole image. Due to the perspective effect, we adopt the ResNet-FPN [8], [34] as the backbone to extract multiscale features in order to make the features invariant to the shape scale. The bottom-up pathway is implemented with ResNet-50 [35], and the output of each stage's last residual block is used as the different-scale features. We use the C2, C3, C4, C5 in [34], which are last residual blocks outputs of conv2, conv3, conv4, conv5 in ResNet. They are with different scales derived from their strides of 4, 8, 16, 32 pixels with respect to the input image. The conv1 is not included due to its large memory footprint.

B. Road Marker Segmentation

We formulate the diamond road marker segmentation as the instance segmentation problem and adopt the Mask-RCNN architecture [36] to segment the individual road marker from input images. It contains three modules: a RPN, a bounding box estimation branch, and a mask estimation branch, as shown in Fig. 2. Different scale features C2, C3, C4, C5 are inputs of RPN for proposal generation. By ROI alignment, features related to each proposal are gathered and input into bounding-box estimation and mask estimation branches. For details about Mask-R CNN, please refer to [36].

C. Lane Segmentation

We design the lane segmentation module based on SCNN network [9]. Spatial convolution layers help propagate the spatial information to the top hidden layer for lane segmentation. Based on our experimental results in Section IX-B, we adopt the

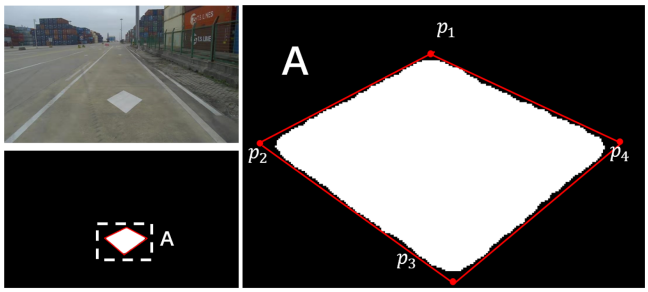


Fig. 3. (a) Raw image on the top and its segmentation mask below. (b) The details in region A. We restore edge intersections as p_1, p_2, p_3, and p_4 for matching with the map. (a) Image and its mask. (b) Refinement for intersections in Region A.

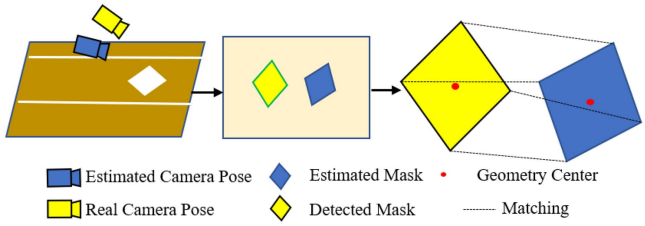


Fig. 4. Illustration of the mask matching with the prebuild map. The blue mask is the bridge to connect the 3-D position from the map with the corresponding detected intersection.

feature C3 from ResNet-FPN as the input feature for the lane segmentation module to balance the accuracy and the memory footprint. Compared to SCNN network [9], a transpose 2-D convolution layer is added before the final convolution layer to improve the resolution of the final lane segmentation results. Besides estimating masks of lanes $\hat{\mathbf{y}}_{\text{LD,mask}} \in \mathbb{R}^{W\times H\times 5}$,[2] the lane segmentation module estimates the existence of lanes as a 4-length tensor $\hat{\mathbf{y}}_{\text{LD,exist}} \in \mathbb{R}^4$ with high-level features.

D. Loss Function

We linearly combine loss functions of the lane segmentation and the road marker segmentation to jointly train two tasks. The loss function $\mathcal{L}$ is defined as

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{SD}} + \mathcal{L}_{\text{LD}}\\ \mathcal{L}_{\text{SD}} &= \mathcal{L}_{\text{SD,cls}} + \mathcal{L}_{\text{SD,box}} + \mathcal{L}_{\text{SD,mask}}\\ \mathcal{L}_{\text{LD}} &= \epsilon_{\text{CE}}(\hat{\mathbf{y}}_{\text{LD,seg}}, \mathbf{y}_{\text{LD,seg}})\\ &\quad + \lambda\epsilon_{\text{CE}}(\hat{\mathbf{y}}_{\text{LD,exist}}, \mathbf{y}_{\text{LD,exist}}),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{SD}}$ and $\mathcal{L}_{\text{LD}}$ denote the loss function for lane segmentation and diamond road marker segmentation. Definitions of $\mathcal{L}_{\text{SD, cls}}$, $\mathcal{L}_{\text{SD, box}}$, $\mathcal{L}_{\text{SD, mask}}$ are the same as [8].

VI. Measurement Postprocessing

We implement two steps to postprocess segmentation masks. We first refine edge intersections from unsmooth masks and create correct correspondences for each feature pair. Then, we determine the heading angle in $\mathbf{F}_W$ by associating lane detection masks with the prebuild map.

A. Segmentation Results Postprocessing

1) Edge Intersection Refinement: Roadmarker segmentation results from Section IV-B are binary 0-1 classification labels to distinguish foreground and background, as shown in Fig. 3(a). Edges of detected roadmarkers are not exactly smooth and accurate as expected due to various reasons such as image resolution, network detection accuracy, and quality of training dataset. Edge Points (whose neighbors' labels with both 0 and 1) on the same edge are divided into groups. For each point k in a group, a line is optimized as

$$\begin{aligned}&\arg\min_{A,B,C}\left\{\frac{1}{N}\sum_{x_i,y_i\in K}\frac{|Ax_i+By_i+C|}{\sqrt{A^2+B^2}}\right\}\\ &s.t.\sum_{x_i,y_i\in K}\frac{|Ax_i+By_i+C|}{\sqrt{A^2+B^2}} < d\end{aligned} \tag{4}$$

where A, B, C are parameters of a line function, K is the point set, with N points, inside a circle with k as the center, and l is the radius. d is the distance threshold for selecting proper lines. Both l and d are tuning parameters. We treat the line as a proper fitting line for one edge if its parameters A, B, C satisfy (4). Four lines are optimized to surround the detected mask as large as possible, and four intersections are generated as the final 2-D points for matching [see Fig. 3(b)].

2) False Detection Rejection: Due to the performance of LDS-Net and challenging working conditions, road marker and lane detection have some false detection. We design a rejection step to filter false detections. With the prior information of road markers and the camera pose, we reproject detected road markers to the 3-D coordinates in $\mathbf{F}_W$ to compare with real road marker's dimension. We also utilize the parallel characteristics with fixed road width for false detection rejection.

3) Feature Correspondence With Prebuild Map: Based on history localization results and prebuild map, the next road markers' 2-D position in the image plane can be estimated. We match estimated and detected roadmarkers with a point-to-point correspondence by aligning their geometry centers and searching the closest point pairs with KD-tree, as shown in Fig. 4. Here, we only match roadmarkers with entire four edge intersections detected.

B. Heading Angle Estimation:

With the help of the homography matrix, we can transfer the lane segmentation results from Section V-C to a bird's-eye-view perspective. We compute the homography matrix by selecting four points in $\mathbf{F}_W$ and corresponding points in the image, which could be edge intersections of a roadmarker. As the large size of the UPV, we adopt both front and rear images to splice a complete lane image. The heading angle in $\mathbf{F}_b$ is the intersection angle

[2] W and H denote the shape of the image, and five denotes the number of classes (the background and four lanes).

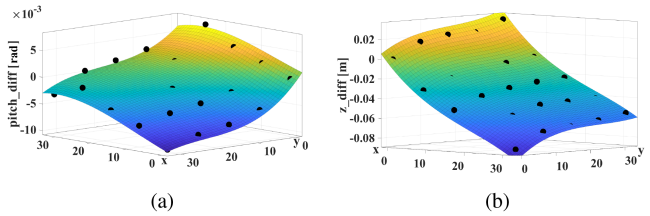


Fig. 5. Illustration of fitting planes for pitch angles (a) and z translations (b). The *x*–*y* axis means the weight of loadings (unit in tons) in the front and rear parts of the UPV. Black dots are data samples. The color of planes varies by the *z*-axis.

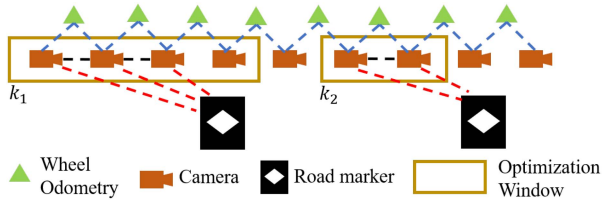


Fig. 6. Illustration of the graph-based optimization model of camera pose estimation. k_1 and k_2 indicate different timestamp. Red, black, and blue lines indicate constrains of PnP matching, feature matching, and wheel odometry, respectively.

between the lane and image vertical central axis. By aligning the detected lane with the prebuild map, the heading angle in $\mathbf{F}_W$ is retrieved. With this information, we refine the wheel odometry's rotation by methods in [37].

VII. Extrinsics Precalibration

Extrinsic parameters are hugely changed with loading variation because of containers' tons of weights. However, existing online approaches only focus on calibration problems caused by mechanical vibration, which are relatively small. We introduce our extrinsic precalibration approach with a structural model as an initial value and refinement with visual constraints. Extrinsic will also be optimized in graph-based optimization in the next section.

A. Polynomial Fitting With Vehicle Structure Model

We describe extrinsic parameters of cameras to $\mathbf{F}_b$ as $\mathbf{T}_{c^i}^b$, where i indicates camera index. We aim to module the $\mathbf{T}_{c^i}^b$ caused by containers' weights and consider this as a polynomial fitting problem. Thanks to the electronic total station, we record the $\mathbf{t}_{c^i}^b$ of each monocular with different loading weights (every 8 tons from 0 to 64 tons) and calculate the corresponding $\mathbf{R}_{c^i}^b$ by solving the 3D–2D [38] homography problem with fixed translation constrains. Third-order fitting is adopted because of nonsensitive to noise observation and less fitting data requirements. We show the fitting plane results with estimated extrinsic parameters in Fig. 5. Loadings mainly affect the pitch and z-axis of the camera extrinsic parameters by analyzing $\mathbf{t}_{c^i}^b$ with corresponding $\mathbf{R}_{c^i}^b$, and we can see that variations in the z-axis are as large as 10 cm. This fitting model provides rough calibration results with the container's weight information as input.

B. Refinement With Epipolar Constrains

The proposed vehicle model can only provide a rough extrinsic because of the uncertainty of vehicles structure consistency. We also adopt epipolar constraints to refine rotation part of $\mathbf{T}_{c^i}^b$. The BRIEF descriptors [39] are utilized to extract visual features, and the fundamental matrix with RANSAC is used for 2D–2D matching outlier rejections. We fix the $\mathbf{t}_{c^i}^b$ and solve a better $\mathbf{R}_{c^i}^b$ as final extrinsic precalibration results.

For each time that UPVs are stationary for container loading or unloading, we will update extrinsics with these precalibration steps. With these extrinsic values as priors, we develop an online optimization module for extrinsic. We treat extrinsics as parameters of the state vector and construct a graph-based optimization module to provide an accurate extrinsic estimation. We introduce this part in the next section.

VIII. Graph-Based Optimization

We formulate the localization problem through an optimization module. This process is inspired by the graph-based optimization approach with a sliding window. The wheel odometry refined by the heading angle estimation provides a high frequency but easy-to-drift pose estimation. Continuous images with the same detected road marker will be a keyframes set to eliminate the drift in the optimization part. We treat image frames with matched roadmarkers as keyframes and others as normal frames.

A. Formulation

The state vector is defined as

$$\begin{aligned}\mathcal{X} &= [\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{c^1}^b, \ldots, \mathbf{x}_{c^i}^b] \\ \mathbf{x}_k &= [\mathbf{q}_{b_k}^W, \mathbf{t}_{b_k}^W], k \in [1, n] \\ \mathbf{x}_{c^i}^b &= [\mathbf{q}_{c^i}^b, \mathbf{t}_{c^i}^b], i \in [1, I]\end{aligned} \tag{5}$$

where $\mathbf{x}_k$ is the pose of the UPV in $\mathbf{F}_W$ at different timestamps, $\mathbf{x}_{c^i}^b$ is the extrinsics from camera i to $\mathbf{F}_b$, I is the number of camera, and n is the size of the window. We only choose state vectors with detected road marker masks into the window for optimization. To estimate correspondences between these states, we build a graph-based optimization module. As shown in Fig. 6, a graphical model is constructed with edge constraints from three approaches: the wheel odometry between two camera poses, visual feature epipolar constraints, and the PnP matching constraints between keyframes and the prebuild map. The wheel odometry provides the motion observation of UPVs in $\mathbf{F}_b$, which can be treated as the IMU observation in the popular visual-inertial-system (VINS) [40], [41]. Feature epipolar constraints are only utilized to match features of detected road markers corners. As proposed in Section I, normal features are not robust in such environments, and we choose the stable features detected by the proposed LDS-Net. We construct an MLE estimation with residual errors as

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} \{f(\mathcal{X})_\alpha + f(\mathcal{X})_\theta + f(\mathcal{X})_\phi\} \tag{6}$$

where $f()_\alpha$ is the error function between keyframes of epipolar constraints, $f()_\theta$ is the reproject error function between keyframes and road markers in $\mathbf{F}_W$, $f()_\phi$ is the wheel odometry preintegration, which we employ from [42], I is an indicator of masks' corners in keyframes, and K is the indicator of road marker corners in $\mathbf{F}$. Our approach outputs optimization results when the window contains road marker detected pose and refined wheel odometry preintegration results in other conditions. Ceres Solver [43] is used to solve the graph-based optimization problem of (6).

B. Optimization With a PnP Solver:

We implement a reliable 2D–3D PnP solution since we have the accurate 3-D positions from the pre-build map. In each keyframe, we provide the detailed information of (6) as

$$f(\mathcal{X})_\theta = \arg\min_{\mathcal{X}} \sum_{i\in I}\sum_{j\in J} \frac{1/e_i}{\sum_{k\in I} 1/e_k}(\mathbf{p}_{i,j} - \pi(\hat{\mathbf{P}_{i,j}}))$$
$$\hat{\mathbf{P}_{i,j}} = \sum_{m\in M} K(\mathbf{R}_b^{c^m}(\mathbf{R}_W^b \mathbf{P}_{i,j} + \mathbf{t}_W^b) + \mathbf{t}_b^{c^m}) \quad (7)$$

where I is the number of detected roadmarkers, J is the number of a road marker's edge intersections, M is the number of cameras, K is camera intrinsics, e_i is the depth from the center of the i_{th} roadmarker in $\mathbf{F}_W$ to the camera, $\pi(\cdot)$ is the projection function. As further masks introduce more errors with the same segmentation accuracy, we denote the inverse depth $1/e_i$ as the weight parameter for reprojection errors.

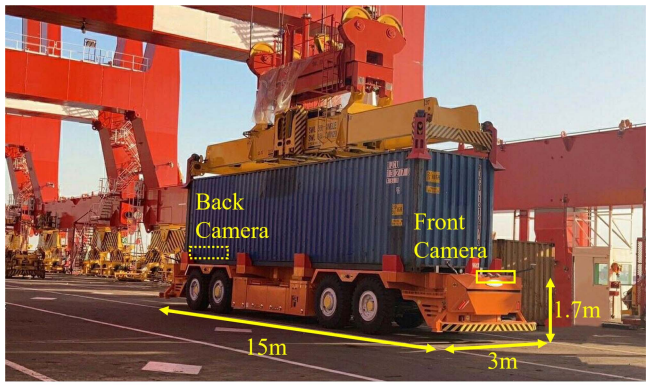
IX. Experiments

We evaluate the segmentation and localization performance of our system with the proposed dataset and real-world SLAM experiments. First, we use the proposed port dataset for roadmarker and lane segmentation, respectively, for the ablation study to compare the performance of the multitask network with a baseline of independent networks. Second, we demonstrate the SLAM performance of the proposed system in port scenarios covering different lighting and speed conditions. We also evaluate the calibration revision with different loading weights by the SLAM performance.

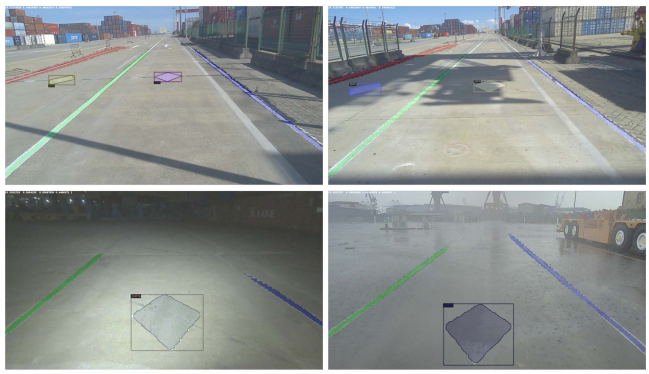
A. Implementation Details

Our dataset evaluation comparisons are conducted on a computer with an Intel i7-7700 K CPU at 4.20 GHz, 32 GB memory, and GeForce 1080 Ti GPU. For outdoor SLAM tasks, we adopt the real-time kinematic (RTK) GPS with 1 Hz frequency as the ground truth. We design experiment trajectories in open areas without tall mechanical structures to ensure GPS accuracy. Monocular cameras capture RGB 752 × 480 images at 20 Hz and are hardware synchronized. Our UPV platform is supported by ZPMC,[3] as shown in Fig. 7(a). For segmentation experiments, we adopt the average precision (AP) as the metric to evaluate the performance of roadmarker segmentation and use the intersection over union (IoU) to evaluate the performance of lane segmentation. SLAM experiments are evaluated by the relative pose error (RPE) and the absolute trajectory error (ATE) [44]. In our system, roadmarkers are set every 15 m, and we use the electronic total station to get road markers' 3-D position.

[3][Online]. Available: https://www.zpmc.com/pro/



(a)



(b)

Fig. 7. We demonstrate the UPV in (a) and the visualization of image segmentation in sunny, shadow, night, and rainy (from left to right and up to down) working scenarios in (b). (a) Brief view of our UPV platform operation. (b) Illustrations of the various challenging scenarios in different weathers.

B. Results on the Proposed Multiobject Dataset

We release our dataset collected in port scenes with lane and roadmarker information in a variety of lighting conditions. We train our LDS-Net on the training set with stochastic gradient descent for 10 epochs. The learning rate is set as 0.01 and decay 30% at epoch 3 and epoch 6.

In this part, our goals include: 1) demonstrate the performance of LDS-Net by jointly solving lane and roadmarker segmentation; 2) discuss the usage of features from ResNet-FPN; 3) discuss the usage of the transposed convolution layer. To achieve these, we compare the following cases.

1) LDS-Net, which is the network detailed in Section V. The C3 feature from ResNet-FPN is adopted as the input of the lane segmentation branch. The transpose convolution layer is adopted to improve the resolution.
2) LS, which contains the ResNet-FPN and lane segmentation branch of LDS-Net.
3) DS, which contains the ResNet-FPN and roadmarker segmentation branch of LDS-Net.

TABLE II
ROAD MARKER SEGMENTATION ON PROPOSED DATASETS

Case	AP	AP50	AP75	APs	APm	APl	time(ms)
LDS-Net	**84.32**	93.23	**93.23**	63.14	**92.13**	**91.53**	99.3
DS	83.12	**94.20**	92.63	**63.18**	89.96	88.58	43.8

AP50, AP75 indicate the IoU as 0.5 and 0.75. APs, APm, and APl mean AP for small, medium, and large objects with 32x32 and 96x96 pixels as boundaries, respectively.

TABLE III
LANE SEGMENTATION ON PROPOSED DATASETS

Case	mIoU	fwIoU	mACC	pACC	time(ms)
LDS-Net	**76.93**	98.16	89.26	**98.98**	99.30
LS-C2	76.92	**98.18**	**90.88**	**98.98**	148.28
LS-C2T	24.28	94.26	25.00	97.05	153.30
LS-C3	71.34	97.54	88.69	98.57	78.40
LS	76.72	98.12	89.39	98.95	79.40
LS-C4	61.77	96.37	83.36	97.76	49.10
LS-C4T	74.00	97.84	88.50	98.77	50.00
LS-C5	48.16	94.33	71.01	96.23	36.40
LS-C5T	65.20	96.87	83.96	98.13	36.30

MIoU, fwIoU, mACC, and pACC Indicate Mean IoU, and Frequency Weighted IoU, Mean Accuracy, and Pixel Accuracy

4) LS-C2, LS-C3, LS-C4, and LS-C5, which are similar to LS. The C2/C3/C4/C5 from ResNet-FPN is adopted as the input of the lane segmentation branch, respectively. No transpose convolution layer is used.
5) LS-C2T, LS-C4T, and LS-C5T, which contains the ResNet-FPN and lane segmentation branch of LDS-Net. The C2/C4/C5 from ResNet-FPN is adopted as the input of the lane segmentation branch. The transpose convolution layer is adopted to improve the resolution.

Experimental results are demonstrated in Tables II and III. They show that LDS-Net, comparing with LD and SD, can detect lanes and squares simultaneously and produce correct results without losing the performance of any task. Due to the shared backbone, the computational efficiency of LDS-Net is better than separate networks LS and DS. The inference time of LDS-Net (99.3 ms) is greatly shorter than the sum of LS and DS (79.4 + 43.8 = 123.2 ms).

Comparing the LS-C3, LS-C4, LS-C5, we find that the C3 is the best feature for lane segmentation. We attribute this to the fact that C3 provides the finest resolution than the others, which helps the highly spatial dependent task, like lane segmentation. By comparing three groups of baseline networks: (LS-Net-C3, LS-Net), (LS-Net-C4, LS-Net-C4T), and (LS-Net-C5, LS-C5T), we find adding a transposed convolution layer with stride 3 improves the resolution of final lane segmentation results. The contribution of fine resolution to the lane segmentation accuracy becomes marginal when using the C2 feature, and it costs much more computational time (approximately 2× slower) than using C3 features. In our experiments, we find that if we use a transposed convolution layer with stride 3 and the C2 feature (LS-Net-C2T), the network does not converge well. The large resolution of C2 feature does not need the improvement provided by the transposed convolution. If we apply transposed convolution to C2 feature, it easily makes the training unstable.

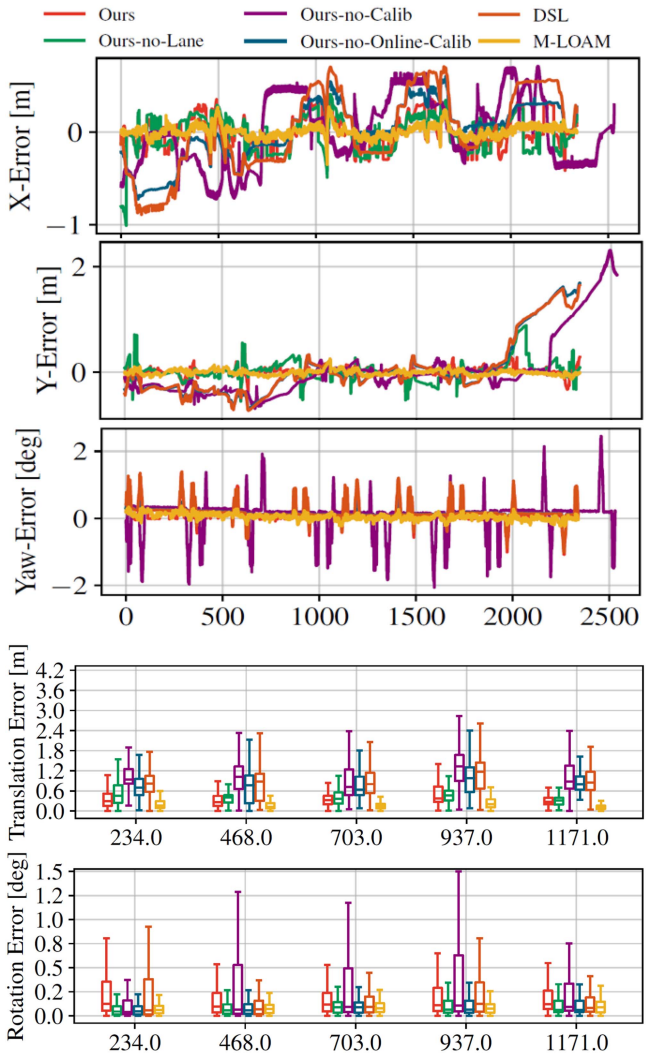


Fig. 8. ATE (top three figures) and RPE (bottom two box figures) in translation and rotation of Section IX-C.

C. Results on Large Scale Outdoor SLAM Experiments

We evaluate our system on an UPV (see Fig. 7) in a port testing field. We compare the proposed system with with two state-of-the-art approaches, visual-based DSL [3] and LiDAR-based M-LOAM [45], as benchmarks. DSL is a monocular visual localization method with a prebuild pointcloud surfel map. M-LOAM is adopted because of its high accuracy for comparison experiments. We also design experiments demonstrating parts of our system performance of how proposed heading angle refinement module and calibration module affect the system accuracy. These are denoted by Ours-no-Lane, Ours-no-Online-Calib, Ours-no-Calib, indicating our methods without heading angle estimation, our method without online calibration, and our methods without precalibration and online calibration. The results are shown in Fig. 8. The GPS is adopted as the ground truth. Only the x–y axis and heading angle are considered as

TABLE IV
ATE RMSE ON SECTION IX-C

Metric	DSL	Ours	Ours-no -Lane	Ours-no -Calib	Ours-no-Online-Calib	M-LOAM
T[m]	0.591	0.178	0.262	0.612	0.558	0.077
R[deg]	0.334	0.158	0.189	0.515	0.199	0.092

TABLE V
ATE RMSE (IN METERS) OF OUR APPROACH, LIDAR-BASED BENCHMARK M-LOAM, AND VISUAL-BASED BENCHMARK DSL OF SEQ.01 TO 07 WITH DIFFERENT SCENES AND LOADINGS

Seq.#	Length	Scene	Loadings	Ours	M-LOAM	DSL
01	2933	Sunny	32 tons	0.122	0.047	0.713
02	1643	Night	16 tons	0.112	0.049	2.582
03	1276	Shadow	24 tons	0.129	0.047	0.733
04	1280	Rainy	0	0.067	0.038	1.581
05	878	Shadow	16 tons	0.105	0.041	0.951
06	1931	Night	64 tons	0.147	0.047	1.958
07	255	Sunny	0	0.081	0.053	0.592

helpful evaluation parameters for experiments. Furthermore, we also design a variety of experiments with different loadings and environments with seven different sequences.

As shown in Fig. 8, Table IV, and Table V, our localization accuracy has better performance than DSL and achieves similar results with the state-of-the-art LiDAR-based method. Even DSL has a reliable prebuild 3-D surfel map, which plays the same role as our prebuild 2-D map, the feature extraction part are easliy affected by appearance changing, leading to drift in featureless environments (night or rainy). The heading angle refinement and two-step calibration module improve the localization accuracy without drift at the end of trajectory. Our method maintains a long-term and high-accuracy localization performance in appearance-changing environments with different loadings. We also notice that the proposed system is hardly affected by scene changes, which proves the segmentation accuracy by the proposed network. However, loading variations affect more than appearance changes. We evaluate the extrinsic perturbation by comparing loadings from 0 to 64 tons, and our ATE increases with loadings. This indicates that the center of gravity location of each container is not the same and fixed. Another reason is that our fitting model for extrinsic parameters has the potential to be improved. For long-term consistency, roadmarkers from the prebuild map provide the loop closure function to ensure drift elimination.

X. DISCUSSION

A. Advantages

We highlight that our proposed system is an accurate and robust visual localization system with online calibration in large-scale challenging environments. A typical application of our system is for autonomous driving of enormous UPVs in ports. Our system complements UGVs' performance, which suffers from localization failure under gantry cranes caused by GPS signals blocked. Our proposed method enables UPVs to extract high-level diamond-shaped road markers as visual observations instead of low-level visual features. As verified by our experiments, the proposed system achieves decimeter-level localization accuracy in various challenging environments with nearly 10 km length of tracks. Compared with other complicated auxiliaries-based [Quick Response (QR) code] visual approaches, we introduce more general diamond-shaped markers for a low cost of construction and maintenance.

B. Limitations

We recognize that our proposed calibration and localization system has limitations. First, our approach relies on a huge number of road markers to be preset and their position information to construct the pre-build map. These constraints limit the application scenarios. This system can only be adopted in limited applications, such as ports. Second, the LDS-Net does not have the ability to decide whether segmentation results are correct or not, and the localization optimization module trusts road marker masks with no doubts. In practice, we observed several bad detection results. To solve this problem, we utilize false detection rejection in Section VI-A and collect more images as training datasets for our LDS-Net. This makes our network overfitting to some extent. Finally, the density of road markers is related to the quality of the wheel odometry since our system provides refined wheel odometry as localization results when there are no road markers available to be detected.

XI. CONCLUSION

In this article, we propose a robust and high-accuracy visual localization pipeline in the large-scale outdoor port scene. Stable visual instance segmentation results are extracted by the proposed LDS-Net for roadmarker segmentation and lane segmentation with real-time performance. The extrinsic variation caused by containers is also considered by vehicle-structure-based model fitting and online optimization. We also propose an optimization module to solve a graph-based optimization problem with refined wheel odometry, keyframe feature matching, and PnP corresponding constraints. Furthermore, we release our image dataset about diamond auxiliaries and lanes in different lighting conditions for segmentation tasks. We demonstrate the robustness and high accuracy performance of the system in real port scenes with a scale of square kilometers.

Our future work will focus on adopting high-level semantic information to improve the robustness of our system. Additionally, we also plan to improve the proposed neural network efficiency and localization accuracy with less computational resource cost. Last but not least, we will investigate general approaches without auxiliaries in similar challenging scenes.

REFERENCES

[1] H. Huang, Y. Sun, H. Ye, and M. Liu, "Metric monocular localization using signed distance field-based maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 1195–1201.
[2] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.
[3] H. Ye, H. Huang, and M. Liu, "Monocular direct sparse localization in a prior 3D surfel map," in *Proc. Int. Conf. Robot. Automat.*, 2020, pp. 8892–8898.

[4] Y. Yu, W. Gao, C. Liu, S. Shen, and M. Liu, "A GPS-aided omnidirectional visual-inertial state estimator in ubiquitous environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 7750–7755.
[5] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2019.
[6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.
[7] R. Fan *et al.*, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE Trans. Mechatronics*, vol. 27, no. 1, pp. 225–233, Feb. 2022.
[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[9] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7276–7283.
[10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[11] W. Cheng, S. Yang, M. Zhou, Z. Liu, Y. Chen, and M. Li, "Road mapping and localization using sparse semantic visual features," *IEEE Robot. Automat. Lett.*, pp. vol. 6, no. 4, pp. 8118–8125, Oct. 2021.
[12] J. Jiao, P. Yun, L. Tai, and M. Liu, "Mlod: Awareness of extrinsic perturbation in multi-lidar 3D object detection for autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10556–10563.
[13] J. Wu *et al.*, "Globally optimal symbolic hand-eye calibration," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 3, pp. 1369–1379, Jun. 2021.
[14] A. Torii, R. Arandjelović, J. Sivic, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1808–1817.
[15] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Heidelberg: Springer, 2010, pp. 255–268.
[16] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE Int. Conf. Robot. Automat.*, 2012, pp. 1643–1649.
[17] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2414–2422.
[18] M. E. Fathy, Q.-H. Tran, M. Z. Zia, P. Vernaza, and M. Chandraker, "Hierarchical metric learning and matching for 2D and 3D geometric correspondences," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 832–850.
[19] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 1651–1662.
[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.
[21] H. Germain, G. Bourmaud, and V. Lepetit, "Sparse-to-dense hypercolumn matching for long-term visual localization," in *Proc. Int. Conf. 3D Vis.*, 2019, pp. 513–523.
[22] H. Huang, H. Ye, Y. Sun, and M. Liu, "Gmmloc: Structure consistent visual localization with Gaussian mixture models," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5043–5050, Oct. 2020.
[23] T. Sattler, A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Are large-scale 3D models really necessary for accurate visual localization?," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6175–6184.
[24] L. Svärm, O. Enqvist, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1455–1461, Jul. 2017.
[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.
[26] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-Roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 340–356.
[27] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.
[28] O. Janssens, R. Van de Walle, M. Loccufier, and S. Van Hoecke, "Deep learning for infrared thermal image based machine health monitoring," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 151–159, Feb. 2018.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
[31] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 75–91.
[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[33] B. Huval *et al.*, "An empirical evaluation of deep learning on highway driving," *CoRR*, 2015.
[34] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[37] C. Belta and V. Kumar, "An SVD-based projection method for interpolation on SE(3)," *IEEE Trans. Robot. Automat.*, vol. 18, no. 3, pp. 334–345, Jun. 2002.
[38] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.
[39] M. Calonder, V. Lepetit, C. Strecha, and F. Brief, "Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 778–792.
[40] S. Tan, S. Zhong, and P. Chirarattananon, "A one-step visual-inertial ego-motion estimation using photometric feedback," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 12–23, Feb. 2022.
[41] L. Jin, H. Zhang, and C. Ye, "Camera intrinsic parameters estimation by visual–inertial odometry for a mobile phone with application to assisted navigation," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 4, pp. 1803–1811, Aug. 2020.
[42] W. Lee, K. Eckenhoff, Y. Yang, and G. Huang, "Visual-inertial-wheel odometry with online calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4559–4566.
[43] S. Agarwal and K. Mierle, "Ceres solver," Mar. 2022. [Online]. Available: https://github.com/ceres-solver/ceres-solver
[44] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 7244–7251.
[45] J. Jiao, H. Ye, Y. Zhu, and M. Liu, "Robust odometry and mapping for multi-LiDAR systems with online extrinsic calibration," *IEEE Trans. Robot.*, vol. 38, no. 1, pp. 351–371, Feb. 2022.

Yang Yu received the B.Eng. degree in electrical engineer from the Donghua University, Shanghai, China, in 2012, and the M.Eng. degreee in electrical engineering from the Illinois Institue of Technology, Chicago, IL, USA, in 2014. He is currently working toward the Ph.D. degree in electrical engineering with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

His research interests include state estimation, SLAM, sensor fusion, and computer vision.

Peng Yun received the B.Sc. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently working toward the Ph.D. degree in electrical engineering with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong.

His current research interests include computer vision, incremental learning, and autonomous driving.

Bohuan Xue (Graduate Student Member, IEEE) received the B.Eng. degree in computer science and technology from College of Mobile Telecommunications, Chongqing University of Posts and and Telecom, Chongqing, China, in 2018. He is currently working toward the Ph.D. degree in electrical engineering with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, HKSAR, China.

His research interests include SLAM, computer vision, and 3D reconstruction.

Jianhao Jiao received the B.Eng. degree in instrument science from Zhejiang University, Hangzhou, China, in 2017. He is currently working toward the Ph.D. degree in electrical engineering with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

His research interests include state estimation, SLAM, sensor fusion, and computer vision.

Rui Fan (Member, IEEE) received the B.Eng. degree in automation from the Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018.

He is currently a (full) Research Professor with the Department of Control Science and Engineering, and Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China.

His research interests include computer vision, machine/deep learning, image/signal processing, autonomous driving, and bioinformatics.

Ming Liu (Senior Member, IEEE) received the B.A. degree in automation from Tongji University, Shanghai, China, in 2005, and the Ph.D. degree in automation from the Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland, in 2013.

He is currently with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, as an Associate Professor. His research interests include appearance-changing environment modeling, deep-learning for robotics, 3-D mapping, machine learning, and visual control.