Commun Transp Res

Research Article

End-to-end Hierarchical Visual Localization with Rasterized and Vectorized HD Map

Jinyu Miao[1,2], Yifei He[1,2], Tuopu Wen[1,2], Kun Jiang[1,2,✉], Kangan Qian[1,2], Zheng Fu[1,2], Yunlong Wang[3], Zhihuang Zhang[4], Weitao Zhou[1,2], Mengmeng Yang[1,2,✉], Jin Huang[1,2], Zhihua Zhong[1,5], Diange Yang[1,2,✉]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640050

ABSTRACT: Accurate localization serves as an important component in autonomous driving systems. Traditional localization methods involve many standalone modules, which require complex hand-crafted rules and costly hyperparameter tuning by trial-and-error, therefore sacrificing the accuracy and generalization. In this paper, we propose an end-to-end visual localization approach, **RAVE**, in which the surrounding images are associated with the HD map data to estimate poses. To ensure high-quality observations for localization, a low-rank flow-based prior fusion module (**FLORA**) is developed to incorporate misaligned map prior into the perceived BEV features. Pursuing a balance among efficiency, interpretability, and accuracy, a hierarchical localization module is proposed, which efficiently estimates poses through a decoupled BEV neural matching-based pose solver (**DEMA**) using rasterized HD map, and then refines the estimation through a Transformer-based pose regressor (**POET**) using vectorized HD map. The experimental results demonstrate that our method can perform robust and accurate localization under varying environmental conditions while running efficiently.

KEYWORDS: Visual Localization, Pose Estimation, End-to-end System, HD Map

1 Introduction

Visual localization plays a necessary role in high-level Autonomous Driving (AD) systems for its ability to provide real-time vehicle poses using an economical sensor suite (Deng et al. 2026). With efforts from worldwide researchers over decades, extraordinary successes have been achieved in this area, improving the accuracy, robustness, and efficiency of visual localization algorithms (Miao et al. 2024).

In the AD society, High-Definition (HD) maps are preferred by resource-limited Intelligent Vehicles (IVs) for they can precisely and compactly represent static driving scenarios as light-weight vectorized semantic map elements (Liu et al. 2020; Wijaya et al. 2024), reducing the requirements of storage space (Xiao et al. 2020; He et al. 2024). High-level IVs commonly perform map matching with HD maps to consistently obtain high-precision poses (Miao et al. 2024). In general, HD map-based visual localization algorithms are designed as a modular pipeline (Xiao et al. 2020, 2018; Wen et al. 2022), where online perception is first performed to perceive semantic elements from images, and then data association is performed between perceived map elements and HD map data, and finally pose estimation can be done based on matching information, as shown in Fig. 1. Although existing solutions have already achieved promising performance in specific scenes, such a modular framework needs complex rules and costly parameter tuning, hindering its large-scale deployment.

On the one hand, traditional visual localization methods rely on precise perception and data association, but cameras inherently lack depth information, which challenges the 2D-3D association between perception results and map data. Some methods project 3D map elements into 2D image planes given pose hypotheses, and perform data association within the 2D image space (Xiao et al. 2018; Liao et al. 2020). However, the perspective effect of the camera model increases the difficulties of data association (Wen et al. 2022). Alternatively, some other methods transform the perceived map elements into the Bird's-Eye-View (BEV) space (Jang et al. 2021; Vivacqua et al. 2017; Deng et al. 2019). While these methods can retain the original shape of map elements, their performance is affected by imprecise calibration and unavoidable vibration of cameras. In this work, we adopt a visual perception and data association approach in BEV space, but leverage a learning-based technique to guarantee the performance. Moreover, to improve the robustness of visual perception against varying environmental conditions, we inject prior knowledge of the driving scenario from maps into perception and specifically address the spatial misalignment problem between online perception and offline HD map.

On the other hand, as a typical state estimation problem, accurate pose estimation remains an open topic for visual localization research. Traditional methods utilize a well-established non-linear optimization-based solver (Xiao et al. 2018, 2020; Wen et al. 2022) or particle filter-based solver (Liao et al. 2020). Zhao et al. tried an Iterative Closest Point (ICP)-based solver to estimate pose by aligning map elements (Zhao et al. 2024). These traditional rule-based pose solvers generally work well, but they all require hyperparameter tuning and rule debugging based on trial-and-error, and they necessitate excellent perception and association, leading to instability when the systems suffer from environmental interference. Moreover, the traditional modular framework bears the risk of information loss and error accumulation along the pipelines (Hu et al. 2023). The input requirements of pose estimation do not guide the design of visual perception and data

[1] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [2] State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [3] Rimian Kaiwu Technology Co., Ltd., Beijing 100083, China. [4] Qcraft Inc., Beijing 100089, China. [5] Chinese Academy of Engineering, Beijing 100088, China.

✉ Corresponding author. E-mail: D. Yang, ydg@mail.tsinghua.edu.cn; K. Jiang, jiangkun@mail.tsinghua.edu.cn; M. Yang, yangmm_qh@mail.tsinghua.edu.cn

Received: April 8, 2026; Revised: Jun 2, 2026; Accepted: August 24, 2026

association yet, resulting in feature misalignment across modules.

Recent advances in end-to-end (E2E) strategies have spurred the development of E2E visual localization algorithms. These learning-based approaches utilize a unified neural network to perceive environmental features, align them with HD maps, and estimate vehicle poses, as shown in Fig. 1, and the model parameters can be automatically tuned in a data-driven manner, thereby eliminating the need for tuning rules and parameters by trial-and-error. These approaches leveraging HD maps follow a similar framework. They extract BEV features from surrounding images and estimate poses by associating these features with local HD map data (ZHANG et al. 2025; He et al. 2024). Trained on extensive datasets, these methods achieve decimeter-level accuracy and show strong robustness across diverse environmental conditions. Nevertheless, existing solutions face critical limitations. They either function as opaque black-box systems (ZHANG et al. 2025; Du et al. 2025) or demand excessive computational resources (He et al. 2024). Striking an optimal balance among efficiency, accuracy, and interpretability remains a key challenge in this emerging field.

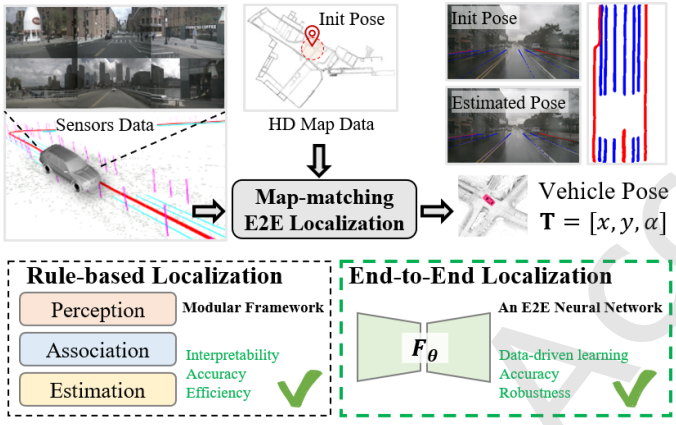

Fig. 1. Traditional rule-based localization methods are built in a modular framework, involving perception, data association, and pose estimation. The end-to-end localization systems directly leverage a unified neural network to estimate the vehicle pose through data-driven model learning, achieving better accuracy and robustness than the traditional methods.

To this end, in this paper, we propose a novel E2E vehicle localization algorithm, **RAVE**, using surrounding images, RAsterized and VEctorized HD maps. In a unified network, the surrounding images are encoded into semantic BEV features, while HD map data are encoded into both rasterized features and vectorized features. In order to obtain stable perceived BEV features for localization, the map prior information is injected into the perceived BEV features by a proposed LOw-RAnk Flow-based prior fusion module (**FLORA**) for online perception enhancement. To achieve efficient and accurate pose estimation, we propose an E2E hierarchical localization module. In the coarse localization stage, a novel DEcoupled BEV neural MAtching-based pose solver (**DEMA**) is proposed in which we can efficiently solve 3 DoF coarse poses by aligning perceived BEV features and rasterized map features in a divide-and-conquer manner. In the fine localization stage, we design a Transformer-based Pose rEgressor (**POET**) to further refine the estimated poses to high precision with fine-grained vectorized map features. The overall network can be E2E optimized, eliminating costly hyperparameter tuning. The primary contributions are summarized as follows:

1) The E2E visual localization network **RAVE** is proposed that performs efficient, accurate, and interpretable pose estimation with lightweight HD maps.

2) The **FLORA** module is proposed to spatially align the perceived BEV features with HD map, and to enhance the online perception by fusing the prior from the offline map.

3) A hierarchical localization module is proposed that utilizes rasterized and vectorized map features. First, the **DEMA** module is proposed to solve 3 DoF poses individually, reducing sampling complexity in neural matching from cubic level to linear level. Second, the **POET** module is designed to refine the estimated poses.

2 Related Works

2.1. Learning-based Online HD Mapping

Recently, constructing HD map elements online through a BEV perception framework has been increasingly adopted by the community as it reduces the substantial human effort needed by conventional rule-based mapping pipelines (Tang et al. 2024; Tang et al. 2025). HDMapNet (Li et al. 2022) builds the first online mapping framework that transforms surrounding images into BEV features by a learnable neural network, fuses cross-modal features and directly learns map elements in BEV space. SuperFusion (Dong et al. 2024) improves the long-range mapping performance by fusing visual and LiDAR data at multiple levels. In contrast to these two methods perceiving rasterized map, VectorMapNet (Liu et al. 2023) directly estimates vectorized HD map elements through a Transformer decoder architecture (Carion et al. 2020). This pipeline facilitates the perception of intricate and detailed shapes of map elements. MapTR series (Liao et al. 2023, 2024) proposes a permutation-equivalent shape modeling method for unified representations of map elements. BeMapNet (Qiao et al. 2023) models semantic map elements as piecewise Bezier curves and achieves excellent perception performance. These works all provide a promising solution for converting raw sensor inputs into BEV features that share the same representation space with HD map, paving the way for data association in subsequent E2E vehicle localization.

2.2. Map Prior-enhanced BEV Perception

Due to occlusion, limited perception field, and varying environmental conditions, achieving stable performance with online perception is theoretically challenging even with sophisticated algorithms. Therefore, many studies have incorporated the prior knowledge from offline map data into online BEV perception (Tang et al. 2024). Some methods learn structural prior of HD map in its offline training stage. P-MapNet (Jiang et al. 2024) utilizes masked autoencoder to learn the distribution of HD map elements so that the imprecise perception can be corrected. DiffMap (Jia et al. 2024), instead, introduces a diffusion module for sampling purposes and formulates the perception enhancement as a denoising process. Some recent methods enhance the mapping performance by incorporating prior maps in the online perception stage, such as Standard Definition (SD) maps (Jiang et al. 2024; Zeng et al. 2024), outdated HD maps (Sun et al. 2025; Zeng et al. 2024), and historically perceived maps (Xiong et al. 2023; Zeng et al. 2024). P-MapNet (Jiang et al. 2024) integrates SD map features using a cross attention mechanism. NMP (Xiong et

al. 2023) builds a global neural map prior during historical inference stage and injects the historical prior by a Gated Recurrent Unit (GRU)-based module in the online perception stage. PriorDrive (Zeng et al. 2024) flexibly integrates different prior maps in both BEV features and learnable queries of the map decoder with a unified framework. In localization tasks where HD maps exist, we can enhance the online mapping performance with offline HD map. However, existing studies ideally assume precise vehicle poses are given. Our work addresses this limitation by considering the spatial alignment between online perception and HD map.

2.3. Visual Localization based on HD Map

Traditional HD map-based localization systems are generally built in a modular framework (Miao et al. 2024). They first extract semantic map elements (*e.g.*, lanes (Pink 2008), poles (Spangenberg et al. 2016), traffic signs (Qu et al. 2015)) from images, then perform map element matching with the HD map, and finally solve poses based on the matching information. However, indistinguishable elements, missed and false detections (Jiang et al. 2023), and inaccurate view transformations (Wan et al. 2024) make the rule-based map matching process highly ambiguous. Moreover, traditional pose solvers including both non-linear optimizers (Xiao et al. 2018; Wan et al. 2024; Wen et al. 2022; Xiao et al. 2020) and particle filters (Liao et al. 2020) heavily rely on precise matching information. Low-quality data association results will result in incorrect convergence in pose estimation (Jiang et al. 2023). Such a modular framework requires careful design of information transfer across modules and needs costly hyperparameter tuning by trial-and-error approaches, blocking the wide usage in varying driving scenarios.

Recently, researchers have begun to develop learning-based localization methods using HD map data. Zhang et al. employed DNN models to directly estimate relative longitudinal and latitudinal positions between the local semantic map and the HD map (Zhang et al. 2022), but the solution requires multiple sensor fusion to build the local semantic map. Zhao et al. utilized a BEV perception network to obtain BEV map elements and estimated poses by a traditional optimization-based solver (Zhao et al. 2024). In pursuit of a fully E2E system that directly estimates vehicle poses from raw sensory data, BEVLocator (ZHANG et al. 2025) employs a Transformer-based network to jointly perform feature extraction, data association, and pose estimation. It achieves decimeter-level localization accuracy in longitudinal and lateral position in challenging benchmarks. Subsequently, EgoVM (He et al. 2024) enhances localization accuracy by using more informative HD maps and a neural matching-based pose solver. OrienterNet (Sarlin et al. 2023) also utilizes neural matching to estimate pose by matching SD maps with monocular image in BEV space. SegLocNet (Zhou et al. 2025) achieves further improvements by fusing surrounding images and LiDAR. All of these works first transform image features into BEV space, then estimate poses based on BEV features and map features by learnable network. They optimize the whole network in an E2E data-driven manner, reducing the information loss and feature misalignment across modules, which motivates our work. However, these solutions are difficult to achieve a balance among accuracy, efficiency, and interpretability. The Transformer-based methods function as black-box systems (ZHANG et al. 2025; Du et al. 2025), while traditional neural matching-based approaches require excessive computational resources (He et al. 2024; Zhou et al. 2025). Besides, existing works have not paid attention to the stability of visual perception, which prevents them from estimating poses reliably under adverse environmental conditions. Our work not only aims to provide stable visual observations for E2E localization, but also achieves interpretable localization that harmonizes efficiency and accuracy.

3 Methodology

In this section, we first define the visual-to-HD map localization problem using surrounding cameras. Next, we present the overall framework of the proposed E2E localization network. Then, we describe the four core modules in the network, i.e., BEV perception backbone, HD map encoder, prior fusion module, and localization module. Finally, the training process and an adaptive localization scheme are given.

3.1. Problem Formulation

Following a common assumption in the AD community that the ground vehicle is attached to the flat ground, the considered vehicle pose in the global frame can be degraded to 3 DoF in BEV space, i.e., $\mathbf{T} = [x, y, \alpha] \in \mathbb{SE}(2)$, where x and y denote the longitudinal and lateral distance between the origins of the ego frame and the global frame, and α is the heading/yaw angle. Generally, a HD map is composed of vectorized semantic map elements, $\boldsymbol{\mathcal{M}} = \{\mathbf{m}_i\}$, where a map element (such as lane divider and pedestrian crossing) can be represented by a set of 2D points in the global frame with semantic category.

Given an initial vehicle pose $\mathbf{T}_0$ that is commonly obtained by GNSS signals, the HD map utilized in visual-to-HD map localization tasks is transformed to a virtual viewpoint on the initial vehicle pose and then cropped to a local map $\boldsymbol{\mathcal{M}}_\mathbf{0} = \mathrm{crop}\big(\mathrm{transform}(\boldsymbol{\mathcal{M}}, \mathbf{T}_0)\big)$. Therefore, the visual-to-HD map localization aims to estimate an optimal vehicle pose $\triangle\,\hat{\mathbf{T}} = [\triangle\,\hat{x}, \triangle\,\hat{y}, \triangle\,\hat{\alpha}] \in \mathbb{SE}(2)$, which aligns the surrounding images $\boldsymbol{\mathcal{I}}$ to the local HD map $\boldsymbol{\mathcal{M}}_\mathbf{0}$:

$$\triangle\,\hat{\mathbf{T}} = \underset{\triangle\mathbf{T}}{\mathrm{argmin}}\; d\big(\boldsymbol{\mathcal{I}}, \mathrm{transform}(\boldsymbol{\mathcal{M}}_\mathbf{0}, \triangle\,\mathbf{T})\big) \tag{1}$$

where $d(\cdot,\cdot)$ measures the differences between cross-modal data. After that, the vehicle pose can be obtained by a pose transformation process:

$$\begin{aligned} \hat{x} &= x_0 + \triangle\,\hat{x} \cdot \cos\alpha_0 - \triangle\,\hat{y} \cdot \sin\alpha_0 \\ \hat{y} &= y_0 + \triangle\,\hat{x} \cdot \sin\alpha_0 + \triangle\,\hat{y} \cdot \cos\alpha_0 \\ \hat{\alpha} &= \alpha_0 + \triangle\,\hat{\alpha} \end{aligned} \tag{2}$$

3.2. Overview of the E2E Localization Network

To perform HD map matching-based localization in an E2E manner, the proposed method incorporates both BEV perception and pose estimation in a unified network. As shown in Fig. 2, the network is fed by the surrounding images $\boldsymbol{\mathcal{I}}$ and the local HD map $\boldsymbol{\mathcal{M}}_\mathbf{0}$. We adopt an online mapping network as the visual backbone to obtain BEV features $\mathbf{F}_\mathcal{I}$ from surrounding images. Meanwhile, the local HD map is represented both by a rasterized semantic mask $\mathbf{M}$ in the BEV space and by a set of vectorized map embeddings $\mathbf{E}$. We extract rasterized map features $\mathbf{F}^R_\mathcal{M}$ from $\mathbf{M}$ by a rasterized map encoder network and vectorized map features $\mathbf{F}^V_\mathcal{M}$ from $\mathbf{E}$ through a vectorized map encoder, respectively. Then, $\mathbf{F}^R_\mathcal{M}$ are utilized to enhance the perception quality of $\mathbf{F}_\mathcal{I}$ as a strong prior via the proposed **FLORA** module. Finally, a hierarchical end-to-end localization module is applied to accurately and efficiently estimate vehicle pose $\triangle\,\hat{\mathbf{T}}_f$, which first estimates coarse pose $\triangle\,\hat{\mathbf{T}}_c$ by utilizing $\mathbf{F}^R_\mathcal{M}$ in the proposed **DEMA** solver, and then refines the estimation with $\mathbf{F}^V_\mathcal{M}$ in the proposed **POET** module.

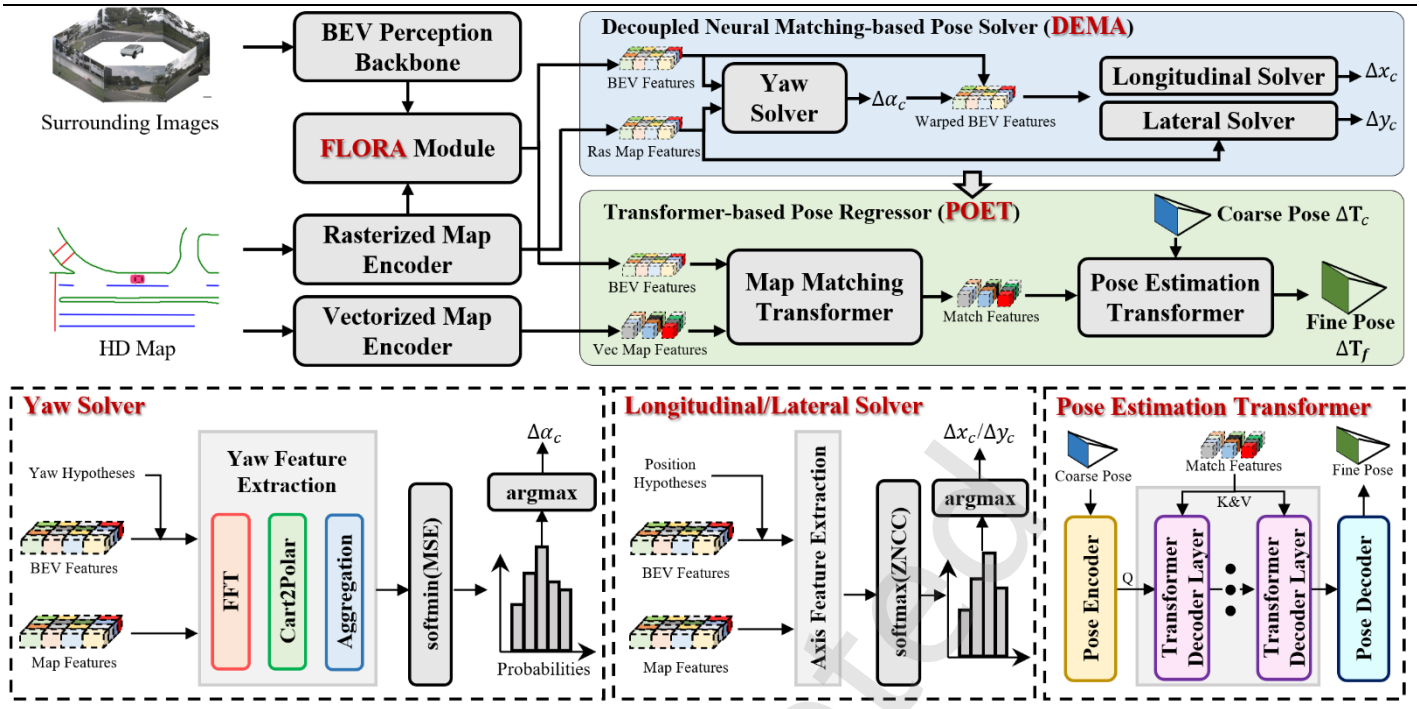

Fig. 2. The overall framework of the proposed end-to-end localization network **RAVE**.

3.3. BEV Perception Backbone and HD Map Encoder

The surrounding images and HD map data are in different modalities, so precise cross-modal matching and pose estimation between them are technically challenging for traditional localization systems. To enable cross-modal matching, we extract high-dimensional learnable features from both surrounding images and HD map data.

First, we employ an online mapping network, BeMapNet (Qiao et al. 2023), as the BEV perception backbone. The backbone takes a set of surrounding images $\boldsymbol{\mathcal{I}}$, intrinsics $\boldsymbol{\mathcal{K}}$ and extrinsics $\boldsymbol{\mathcal{O}}$ of cameras as input, and extracts Perspective-View (PV) features from each image. Then, the backbone transforms and aggregates PV features into BEV space by a Transformer-based view transformation module, and finally obtains BEV semantic features $\mathbf{F}_{\mathcal{I}}$. We use a simple convolution segmentation head $\mathcal{H}$ to decode $\mathbf{F}_{\mathcal{I}}$ to a semantic mask $\widehat{\mathbf{M}}_{\mathcal{I}} \in (0,1)^{H\times W\times(C+1)}$, where H, W are the sizes of the map mask, and C is the number of semantic categories:

$$\mathbf{F}_{\mathcal{I}} = \text{imgbackbone}(\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{O}}), \widehat{\mathbf{M}}_{\mathcal{I}} = \mathcal{H}(\mathbf{F}_{\mathcal{I}}) \quad (3)$$

Since the map elements are sparse in the HD map, we add a background class in the semantic mask to make the BEV features informative for subsequent E2E localization.

Meanwhile, we build a VGG-like network (Simonyan and Zisserman 2015) as the rasterized map encoder to extract rasterized map features $\mathbf{F}^R_{\mathcal{M}}$ from rasterized HD map mask $\mathbf{M} \in \{0,1\}^{H\times W\times(C+1)}$. We reconstruct the map mask from $\mathbf{F}^R_{\mathcal{M}}$ so that the map features can maintain semantic and spatial information of the HD map.

$$\mathbf{F}^R_{\mathcal{M}} = \text{encoder}_{\text{ras}}(\mathbf{M}), \widehat{\mathbf{M}}_{\mathcal{M}} = \mathcal{H}(\mathbf{F}^R_{\mathcal{M}}) \quad (4)$$

where $\widehat{\mathbf{M}}_{\mathcal{M}}$ is the reconstructed HD map mask.

Additionally, we extract fine-grained map features from vectorized HD map data. Concretely, the HD map data is represented by element-wise hybrid embeddings. For each map element $\mathbf{m}_i$, the 2D positions and directions of its map points are embedded by Fourier position encoding (Tancik et al. 2020):

$$\mathbf{E}_i^{pos} = \text{Fourier}([u_i^0, u_i^1, \dots]), \mathbf{E}_i^{dir} = \text{Fourier}([v_i^0, v_i^1, \dots]) \quad (5)$$

where u_i^k is the 2D position of the k-th point in the $\mathbf{m}_i$ and $v_i^k = u_i^{k+1} - u_i^k$ is the direction between adjacent points in the $\mathbf{m}_i$. The semantic category c_i and the index of the element i are encoded by learnable embeddings, $\mathbf{E}_i^{sem}$, $\mathbf{E}_i^{ins}$, respectively. Thus, a map element is encoded as a fixed-size structured embedding by concatenating all the map information:

$$\mathbf{E}_i = \left[\mathbf{E}_i^{pos}, \mathbf{E}_i^{dir}, \mathbf{E}_i^{sem}, \mathbf{E}_i^{ins}\right] \in \mathbb{R}^{N_p\times D} \quad (6)$$

where N_P is the desirable number of map points in each map element and D is the feature dimension. For those elements without sufficient map points, we pad the elements with zeros. Finally, the vectorized map features $\mathbf{F}^V_{\mathcal{M}}$ are obtained by aggregating map embeddings within each map element through a vectorized encoder (Gao et al. 2020):

$$\mathbf{F}^V_{\mathcal{M}} = \text{encoder}_{\text{vec}}(\mathbf{E}) \quad (7)$$

By converting raw images and HD map data into learnable features, the difficulties of cross-modal data association, fusion, and pose estimation can be effectively addressed by E2E learning.

3.4. FLORA: Low-rank Flow-based Prior Fusion Module

As the visual measurements input to the E2E localization framework, the BEV features from surrounding cameras play an important role in localization stability and accuracy. In this work, the BEV features are trained to encode static map element information, which share the same feature space with HD map data. Thus, the offline HD map will be a strong prior to enhance online BEV perception (Jia et al. 2024; Jiang et al. 2024). However, without high-precision vehicle pose, there is a spatial displacement problem between BEV perception results and the retrieved local HD map, degrading the performance of BEV perception enhancement. Therefore, we propose a flow-based prior fusion module, **FLORA**, which first corrects the spatial displacements between BEV features and rasterized map features based on estimated flows, and then fuses them to enhance BEV perception.

In order to estimate the pixel-level displacement between $\mathbf{F}_{\mathcal{I}}$ and $\mathbf{F}^R_{\mathcal{M}}$, we build a light-weight flow estimation network, as shown in

Fig. 3. Following (Teed and Deng 2020), a shared feature encoder is utilized to extract features from both $\mathbf{F}_{\mathcal{I}}$ and $\mathbf{F}_{\mathcal{M}}^{R}$, and then a multi-scale local correlation volume is obtained by calculating the pixel-wise similarities between neighboring pixels across two features. Since $\mathbf{F}_{\mathcal{M}}^{R}$ is more stable and accurate, it is selected to generate the initial hidden states and the context feature through a context encoder. Under an iterative refinement framework, the network retrieves corresponding correlation data from correlation volume based on estimated flow from the previous iteration, and updates the estimated flow through an update module. In the update module, the correlation data, previously estimated flow, context feature, and previous hidden states are aggregated and updated, resulting in a flow feature, which is finally decoded to BEV flow.

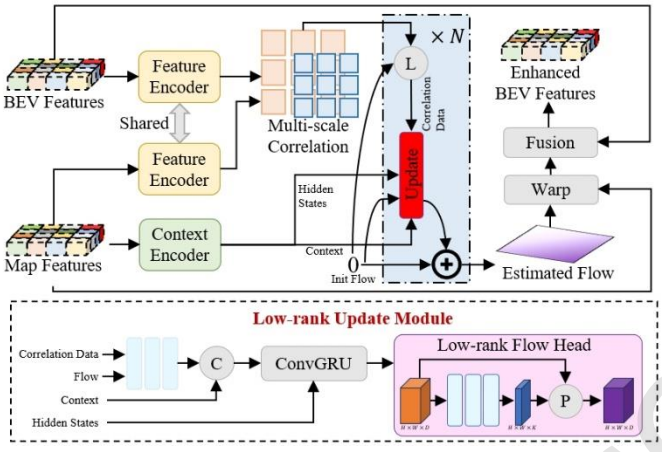

Fig. 3. The structure of the proposed prior fusion network **FLORA**.

Low-rank BEV Flow: In the proposed **FLORA** module, the update module involves a low-rank feature projection process. Given that the spatial displacement between BEV features and map data is caused by 3 DoF pose errors, the displacement in the BEV space should be low-rank and regular. Thus, inspired by (Ye et al. 2021), we reduce the rank of the flow feature so that the dominant flow feature can be retained while noise will be rejected. Formally, the flow feature $\mathbf{F}_f \in \mathbb{R}^{H_f \times W_f \times D}$ is compressed in the channel dimension to reduce its rank, resulting in a feature $\mathbf{F}_v \in \mathbb{R}^{H_f \times W_f \times K}$ with K channels. Each channel of $\mathbf{F}_v$ can serve as a flow feature basis $v_k \in \mathbb{R}^{H_f W_f}, k = 1,2,\ldots,K$ after flattening. Finally, the original flow feature $\mathbf{F}_f$ is projected into the subspace of the flow feature bases so that we obtain a low-rank flow feature $\mathbf{F}_f^l \in \mathbb{R}^{H_f \times W_f \times D}$:

$$\mathbf{F}_f^l = V(V^T V)^{-1} V^T \cdot \mathbf{F}_f \tag{8}$$

where $V = [v_1, \ldots, v_K]$. The normalization term $(V^T V)^{-1}$ is required since v_k is not guaranteed to be orthogonal.

With the estimated flow from the **FLORA** module, the rasterized map feature $\mathbf{F}_{\mathcal{M}}^{R}$ is warped so that the spatial displacement is effectively reduced. The warped map feature and the BEV feature are spatially aligned. Therefore, we simply concatenate these two features in channel dimension and fuse them by a convolution layer. The overall procedure in the proposed **FLORA** module can be formulated as:

$$\mathbf{F}_{\mathcal{I}}^{e} = \mathbf{FLORA}(\mathbf{F}_{\mathcal{I}}, \mathbf{F}_{\mathcal{M}}^{R}) = \mathrm{conv}([\mathbf{F}_{\mathcal{I}}, \mathrm{flow}(\mathbf{F}_{\mathcal{I}}, \mathbf{F}_{\mathcal{M}}^{R}) \otimes \mathbf{F}_{\mathcal{M}}^{R}]) \tag{9}$$

where $\otimes$ denotes the warp operator based on estimated flow.

By fusing strong prior information from the HD map features, the quality of BEV perception feature is significantly improved. We also decode the enhanced BEV features $\mathbf{F}_{\mathcal{I}}^{e}$ to a map mask for training: $\hat{\mathbf{M}}_{\mathcal{I}}^{e} = \mathcal{H}(\mathbf{F}_{\mathcal{I}}^{e})$.

3.5. Hierarchical End-to-end Localization Module

After obtaining enhanced high-quality BEV features $\mathbf{F}_{\mathcal{I}}^{e}$, rasterized map features $\mathbf{F}_{\mathcal{M}}^{R}$, and vectorized map features $\mathbf{F}_{\mathcal{M}}^{V}$, we propose to estimate vehicle pose relative to the local HD map $\boldsymbol{\mathcal{M}}_{\mathbf{0}}$ in a coarse-to-fine localization framework, where we utilize rasterized HD map features in the coarse stage and vectorized HD map features in the fine stage. In the coarse stage, we efficiently solve poses by aligning rasterized features, but its localization accuracy is limited by BEV perception resolution. Therefore, in the fine stage, fine-grained vectorized map features are leveraged to refine estimated pose through a deep Transformer-based pose regressor.

3.5.1 Decoupled BEV Neural Matching-based Coarse Localization using Rasterized Map

Considering that the spatial displacement between BEV perception and HD map is caused by inaccurate 3 DoF vehicle pose, we first utilize a decoupled neural matching-based solver **DEMA** for interpretable and efficient coarse pose estimation using $\mathbf{F}_{\mathcal{I}}^{e}$ and $\mathbf{F}_{\mathcal{M}}^{R}$:

$$\triangle \hat{\mathbf{T}}_c = \mathbf{DEMA}(\mathbf{F}_{\mathcal{I}}^{e}, \mathbf{F}_{\mathcal{M}}^{R}) \tag{10}$$

Note that we compress the channel dimension and size of both $\mathbf{F}_{\mathcal{I}}^{e}$ and $\mathbf{F}_{\mathcal{M}}^{R}$ before feeding them into the **DEMA** module for computational efficiency, *i.e.*, $\mathbf{F}_{\mathcal{I}}^{e}, \mathbf{F}_{\mathcal{M}}^{R} \in \mathbb{R}^{H_c \times W_c \times D_c}$.

Full Neural Matching (FUMA): Traditional **FUMA** solver in BEV space exhaustively samples all possible 3 DoF pose hypotheses by grid sampling: $\{\triangle \mathbf{T}_{ijk} = [\triangle x_i, \triangle y_j, \triangle \alpha_k]\}$, where $i \in [1, N_x], j \in [1, N_y], k \in [1, N_\alpha]$ and N_x, N_y, N_α are the numbers of hypotheses sampled in each dimension, and then estimates their probabilities based on the feature similarities (Weston et al. 2022; He et al. 2024; Hosseinzadeh and Reid 2024). Theoretically, the number of pose hypotheses in the **FUMA** solver is as high as $N_x \times N_y \times N_\alpha$, leading to unacceptable computational burden when the hypothesis space is extremely large.

Decoupled Neural Matching (DEMA): In this paper, we propose to decouple the feature representation affected by each DoF in pose estimation and estimate each DoF through the **DEMA** solver in a divide-and-conquer framework. As shown in Fig. 2, the proposed pose solver first extracts yaw features from $\mathbf{F}_{\mathcal{I}}^{e}$ and $\mathbf{F}_{\mathcal{M}}^{R}$ and estimates the yaw angle by measuring differences between yaw features. Second, $\mathbf{F}_{\mathcal{I}}^{e}$ is transformed to $\mathbf{F}_{\mathcal{I}}^{\alpha}$ by the estimated yaw angle $\Delta\hat{\alpha}$ so that $\mathbf{F}_{\mathcal{I}}^{\alpha}$ and $\mathbf{F}_{\mathcal{M}}^{R}$ differ only in longitudinal and lateral positions. Finally, we extract axis features from $\mathbf{F}_{\mathcal{I}}^{\alpha}$ and $\mathbf{F}_{\mathcal{M}}^{R}$ and solve longitudinal and lateral positions. Detailed descriptions are given below.

Yaw Solver by Amplitude-Frequency Features: As shown in Fig. 4, the amplitude-frequency graph of an image is only affected by its orientation but not its position (Weston et al. 2022; Fan et al. 2022). Therefore, the yaw solver of the **DEMA** module individually solves yaw angle by decoupling the BEV feature representation affected by yaw angle. Formally, we generate yaw hypotheses by grid sampling: $\{\triangle \alpha_k | 1 \le k \le N_\alpha\}$ and transform $\mathbf{F}_{\mathcal{I}}^{e}$ accordingly, resulting in transformed BEV features $\{\mathbf{F}_{\mathcal{I}}^{\alpha(k)} | 1 \le k \le N_\alpha\}$. Then, $\mathbf{F}_{\mathcal{I}}^{\alpha(k)}$ and $\mathbf{F}_{\mathcal{M}}^{R}$ are converted into frequency domain by Fast Fourier Transform (FFT) process to obtain Frequency-BEV (F-BEV) features. To quantitatively represent the

amplitude-frequency features affected by yaw angle along the yaw axis in polar coordinates, we transform the F-BEV features in Cartesian coordinates into Polar F-BEV (PF-BEV) features in polar coordinates. Finally, we aggregate the PF-BEV features perpendicular to the yaw axis and obtain corresponding *yaw features*:

$$\mathbf{Y}_* = \mathrm{sum}\left(\mathrm{Cart2Polar}\left(\mathrm{FFT}(\mathbf{F}_*)\right)\right) \tag{11}$$

The probability of each yaw hypothesis is calculated by normalized Mean Square Error (MSE) between *yaw features*:

$$p(\triangle \alpha_k) = \mathrm{softmin}\left(\mathrm{MSE}\left(\mathbf{Y}_{\mathcal{M}}, \mathbf{Y}_J^{\alpha(k)}\right)\right) \tag{12}$$

and an optimal yaw angle $\Delta\hat{\alpha}$ can be estimated:

$$\triangle \hat{\alpha}_c = \triangle \alpha_k, k = \underset{k}{\mathrm{argmax}}\, p(\triangle \alpha_k) \tag{13}$$

The estimated yaw angle is utilized to transform $\mathbf{F}_J^e$, correcting the orientation differences between features.

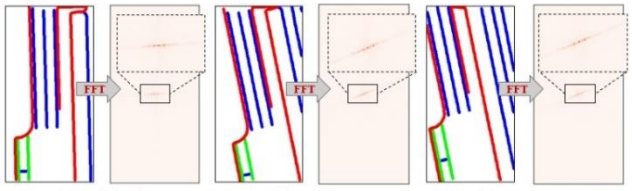

Fig. 4. Visualization cases of the amplitude-frequency features for map mask. The left image in each group is HD map mask while the right one is its amplitude-frequency graph. (a) shows an original HD map, (b) shows a map after adding 10° yaw deviation, and (c) shows a map after adding 10° yaw deviation and 5m deviation in both longitudinal and lateral directions.

Longitudinal/Lateral Solver by Axis Features: The position differences of features in longitudinal dimension can be quantitatively reflected in the row order of features, as shown in Fig. 5 (a) and (c), but the position difference in the other dimension (lateral) will make this phenomenon indistinct. Therefore, to individually solve longitudinal and lateral positions, we propose to aggregate features along the single dimension in the longitudinal and lateral solvers so that the aggregated axis features differ only in the position of one dimension, as shown in Fig. 5 (b) and (d). For instance, to solve longitudinal position Δx, the longitudinal solver samples longitudinal hypotheses by grid sampling: $\{\triangle x_i | 1 \le i \le N_x\}$ and transforms $\mathbf{F}_J^e$ into features $\{\mathbf{F}_J^{x(i)} | 1 \le i \le N_x\}$. The transformed BEV features and map features are aggregated by GeM pooling (Radenovic et al. 2019) along a single row dimension and generate corresponding axis feature $\mathbf{A}_* \in \mathbb{R}^{H_c \times D_c}$:

$$\mathbf{A}_*[u] = \left(\frac{1}{W_c}\sum_{v=1}^{W_c}(\mathbf{F}[u,v])^{\omega}\right)^{\frac{1}{\omega}} \tag{14}$$

where u, v are the row and column indices, ω is the learnable weighting factor. Then, the probability of each longitudinal hypothesis is the normalized Zero-mean Normalized Cross-Correlation (ZNCC) between axis features:

$$p(\triangle x_i) = \mathrm{softmax}\left(\mathrm{ZNCC}\left(\mathbf{A}_{\mathcal{M}}, \mathbf{A}_J^{x(i)}\right)\right) \tag{15}$$

and an optimal longitudinal position $\triangle \hat{x}_c$ is estimated. The lateral solver operates in a technically similar way but it aggregates features along a single column dimension. Finally, we obtain a coarse vehicle pose estimation $\triangle \hat{\mathbf{T}}_c = [\triangle \hat{x}_c, \triangle \hat{y}_c, \triangle \hat{\alpha}_c]$.

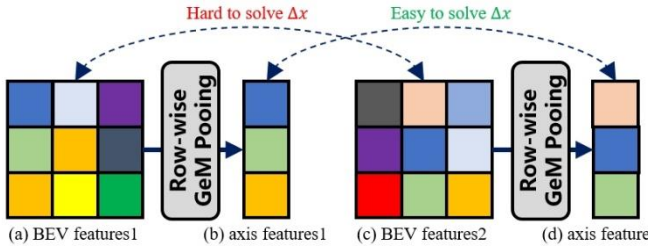

Fig. 5. Diagram of axis feature extraction by GeM pooling (Radenovic et al. 2019). (a) and (c) show two original BEV features whereas (b) and (d) are their corresponding axis features.

3.5.2 Transformer-based Fine Localization using Vectorized Map

Although the **DEMA** solver provides efficient pose estimation in the coarse localization stage, its localization accuracy is still limited, especially in the longitudinal pose estimation. We attribute this limitation to the resolution constraints of BEV perception and map encoding, which restrict the granularity of input features fed into the pose estimation module. The **DEMA** solver operates by aligning rasterized features, therefore, using low-resolution rasterized features impairs its pose estimation accuracy. In order to achieve localization with higher accuracy and correct unreliable pose estimation from the coarse localization module, we propose a Transformer-based pose regressor **POET** to refine pose estimation using enhanced BEV features $\mathbf{F}_J^e$, vectorized map features $\mathbf{F}_{\mathcal{M}}^V$, and coarse pose estimation $\triangle \hat{\mathbf{T}}_c$ with corresponding probability $\hat{\mathbf{P}}_c = \{p(\triangle \hat{x}_c), p(\triangle \hat{y}_c), p(\triangle \hat{\alpha}_c)\}$:

$$\triangle \hat{\mathbf{T}}_f = \mathbf{POET}\left(\mathbf{F}_J^e, \mathbf{F}_{\mathcal{M}}^V \middle| \triangle \hat{\mathbf{T}}_c, \hat{\mathbf{P}}_c\right) \tag{16}$$

In order to make Transformer-based neural network applicable in map matching-based localization tasks, we first initialize the queries for Transformer based on the estimated coarse pose so that the built queries initially contain coarse pose information if the coarse estimation can be referenced. Concretely, the estimated coarse pose probability $\hat{\mathbf{P}}_c$ is converted to high-dimensional pose features via a Multi-layer Perceptron (MLP) module and enhanced by a two-layer Transformer encoder, generating queries $\boldsymbol{Q_T}$ that encode referenced information from the coarse pose estimation, dubbed as *pose queries*:

$$\boldsymbol{Q_T} = \mathrm{selfattn}\left(\mathrm{MLP}\left(\hat{\mathbf{P}}_c\right)\right) \tag{17}$$

Then, as shown in Fig. 2, the enhanced BEV features $\mathbf{F}_J^e$ and vectorized map features $\mathbf{F}_{\mathcal{M}}^V$ are matched using a Transformer module to produce *matching features* $\mathbf{F}_{\mathcal{M}}^M$. This approach avoids generating a huge cost volume (Miao et al. 2023) while efficiently obtaining map matching information:

$$\mathbf{F}_{\mathcal{M}}^M = \mathrm{transformer}(q = \mathbf{F}_{\mathcal{M}}^V, k = v = \mathbf{F}_J^e) \tag{18}$$

The *pose queries* $\boldsymbol{Q_T}$ serve as queries across multiple Transformer decoder layers, enabling iterative interaction with *matching features* (keys and values):

$$\boldsymbol{Q_T} = \mathrm{crossattn}(q = \boldsymbol{Q_T}, k = v = \mathbf{F}_{\mathcal{M}}^M, pos = \boldsymbol{E_T}) \tag{19}$$

In each iteration, the *pose queries* are refined using the relevant information from *matching features*, mimicking the pose optimization process in the traditional map matching-based localization but in an implicit way. Fifteen learnable position embeddings $\boldsymbol{E_T}$ are employed in the Transformer decoder layer, serving as implicit directional guidance for pose optimization as shown in Fig. 6. This allows the network to optimize the pose in

various directions according to different retrieved matching information. After several iterations, the updated *pose queries* encode the implicit pose information between BEV features and map data, which can be decoded to the fine vehicle pose $\triangle \mathbf{T}_f$ with high-precision. Specifically, each updated pose query is decoded into a 3-DoF pose and a corresponding score, and the results are aggregated using normalized scores s.

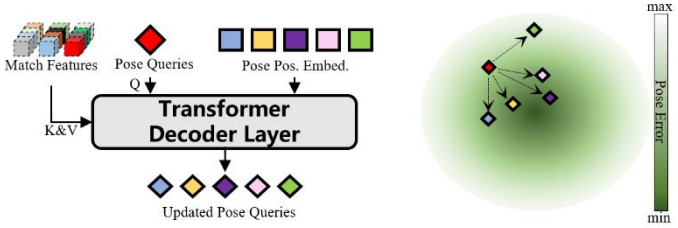

Fig. 6. A diagram of iterative pose optimization process based on Transformer decoder layer in the proposed **POET** module.

3.6. Training Scheme

First, we generate training and validation datasets as in (ZHANG et al. 2025; Du et al. 2025; Yong et al. 2026). We add random noise to real vehicle poses with longitudinal deviation $\triangle x^*$, lateral deviation $\triangle y^*$, and yaw deviation $\triangle \alpha^*$. The HD maps $\mathcal{M}$ are transformed and cropped by the noisy vehicle poses, and fed into the proposed network as $\mathcal{M}_0$.

For stable and fast convergence in network training, we conduct a two-stage training scheme. In the first stage, we train the BEV perception backbone, the rasterized map encoder, and the **FLORA** module through the semantic supervision $\mathcal{L}_{sem}$ (Qiao et al. 2023) on both $\widehat{\mathbf{M}}_j^e$ and $\widehat{\mathbf{M}}_{\mathcal{M}}$. This stage ensures that the enhanced BEV features and the rasterized map features contain implicit semantic map element information, which is matchable for the pose solver. We do not add additional supervision on the estimated flow from the **FLORA** module and only guide it to enhance the BEV features through self-learning flow estimation.

Then, in the second stage, we further add the probability supervision $\mathcal{L}_{prob}$ on the coarse **DEMA** module and the pose supervision $\mathcal{L}_{pose}$ on the fine **POET** module for E2E network training. The localization network aims to estimate the added pose deviations ($\triangle \mathbf{T}^* = [\triangle x^*, \triangle y^*, \triangle \alpha^*]$) based on $\boldsymbol{\mathcal{M}_0}$ and $\boldsymbol{\mathcal{I}}$. Therefore, $\triangle \mathbf{T}^*$ can be utilized as the supervision for pose estimation, allowing the network to be E2E trained.

Concretely, the proposed **DEMA** module provides the probability distribution of 3 DoF pose hypotheses, i.e., $\widehat{\mathbf{P}}_x = \{p(\triangle x_i)\}_{i=1}^{N_x}, \widehat{\mathbf{P}}_y = \{p(\triangle y_j)\}_{j=1}^{N_y}, \widehat{\mathbf{P}}_\alpha = \{p(\triangle \alpha_k)\}_{k=1}^{N_\alpha}$. Generally, we hope the pose hypotheses near the ground-truth pose $\triangle \mathbf{T}^*$ have higher probabilities while decreasing the probabilities of other pose hypotheses. Thus, we add probability supervision $\mathcal{L}_{prob}$ on the probability distributions of 3 DoF pose hypotheses. For example, we sample the probability of ground-truth yaw deviation $\triangle \alpha^*$ from the probability distribution of yaw hypotheses $\widehat{\mathbf{P}}_\alpha$, and calculate its negative logarithmic value as the probability loss:

$$\mathcal{L}_{prob}(\widehat{\mathbf{P}}_\alpha) = -\log\left(\text{sample1d}(\widehat{\boldsymbol{P}}_\alpha, \triangle \alpha^*)\right) \tag{20}$$

As the loss decreases, the probability of $\triangle \alpha^*$ will increase so that the probability distributions from the **DEMA** module will gradually turn into peaky distributions centered on the ground-truth poses as training proceeds. The probability losses for longitudinal and lateral positions are similar to the yaw angle. Because the probability loss supervises the probability distribution estimated by the **DEMA** solver, we can still train its network parameters in an E2E manner, despite the fact that the solver uses non-differentiable argmax operators to derive a coarse pose from the distribution.

As for the fine pose estimation from the proposed **POET** module, we calculate the weighted smooth L1 distance between $\triangle \mathbf{T}_f$ and $\triangle \mathbf{T}^*$ as the pose loss:

$$\mathcal{L}_{pose}(\triangle \mathbf{T}_f) = \rho \cdot \|\triangle \widehat{\mathbf{T}}_f - \triangle \mathbf{T}^*\|_{S1} \tag{21}$$

where ρ is calculated based on the normalized score s:

$$\rho = \frac{\max(s) - \min(s)}{2 \cdot \text{mean}(s)} \tag{22}$$

which can be seen as the reliability of pose aggregation in the **POET** module. We additionally add a reliability loss to encourage an increase in ρ:

$$\mathcal{L}_{rel}(\rho) = -\frac{1}{2}\log(\rho^2) \tag{23}$$

The pose loss and reliability loss are calculated for all estimates from each layer of the **POET** module.

Thus, the final loss in the second E2E training stage is:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{sem}(\widehat{\mathbf{M}}_j^e) + \mathcal{L}_{sem}(\widehat{\mathbf{M}}_{\mathcal{M}}) + \lambda \cdot \left(\mathcal{L}_{pose}(\triangle \widehat{\mathbf{T}}_f) + \mathcal{L}_{rel}(\rho)\right) \\ +0.1 \cdot \left(\mathcal{L}_{prob}(\widehat{\mathbf{P}}_x) + \mathcal{L}_{prob}(\widehat{\mathbf{P}}_y) + \mathcal{L}_{prob}(\widehat{\mathbf{P}}_\alpha)\right)\end{aligned} \tag{24}$$

where λ is an adaptive weight factor based on relative value $r = \mathcal{L}_{pose}(\triangle \mathbf{T}_f)/\mathcal{L}_{sem}(\widehat{\mathbf{M}}_j^e)$:

$$\omega = \begin{cases} 10 & \text{if} \quad r < 0.1 \\ 0.1 & \text{if} \quad r > 10 \\ 1 & \text{otherwise} \end{cases} \tag{25}$$

3.7. Adaptive Coarse-to-fine Localization

Although refining the coarse pose by the Transformer-based **POET** module can effectively improve localization accuracy, it inevitably imposes significantly more computational burden than utilizing the **DEMA** module only. To achieve a balance between localization performance and efficiency, we leverage the estimated probability of the **DEMA** module as an estimated reliability of the coarse pose, adaptively triggering the **POET** module when needed.

Concretely, if the estimated probability (reliability) of the coarse pose is greater than the predefined threshold t_c, we believe the coarse pose is sufficiently reliable and do not refine it through the fine localization module:

$$\triangle \widehat{\mathbf{T}} = \begin{cases} \triangle \widehat{\mathbf{T}}_c & \text{if} \quad \widehat{\mathbf{P}}_c \geq t_c \\ \triangle \widehat{\mathbf{T}}_f & \text{otherwise} \end{cases} \tag{26}$$

By applying such an adaptive strategy, the heavy **POET** module is not activated when the **DEMA** module already estimates vehicle pose reliably, reducing unnecessary computational overhead in the E2E localization pipeline.

4 Experiments

4.1. Implementation Details

Model: We implemented the proposed network using the PyTorch library. The proposed network is trained using AdamW optimizer, a batch size of 8, an initial learning rate of $1e-4$, a weight decay rate of $1e-7$ on 4 NVIDIA RTX 3090 GPUs. The first stage trains for 10

epochs, followed by 15 epochs for full model training in the second stage. Following BeMapNet (Qiao et al. 2023), the BEV perception focuses on three kinds of static map elements, *i.e.*, lane dividers, pedestrian crossings, and road boundaries. The perception ranges cover $[-15.0m, 15.0m]$ laterally and $[-30.0m, 30.0m]$ longitudinally. The resolution of BEV grid is set as 0.15m/pixel, resulting in $H \times W = 400 \times 200$ HD map masks. For fair comparison, the sampling ranges of pose hypotheses are set to $\pm 2m$, $\pm 1m$, $\pm 2^\circ$ in longitudinal, lateral, and yaw dimensions, and the sampling steps are set to $0.4m$, $0.2m$ and 0.4° for efficiency unless otherwise specified. Therefore, the numbers of pose hypotheses in each dimension all equal to 11. The feature dimensions and sizes are set as $H_f \times W_f = 50 \times 25$, $D = 128$, $H_c \times W_c = 100 \times 50$, $D_c = 16$, and $K_f = 16$ by default. We utilize 2 Transformer encoder layers and 2 Transformer decoder layers in the proposed **POET** module. The full model with adaptive strategy can run in less than 150 milliseconds on a commercial NVIDIA RTX 3090 GPU.

Benchmark: The nuScenes dataset (Caesar et al. 2020) is utilized in this paper, which consists of 28,130 training and 6,019 validation samples from 700 training and 150 validation driving scenes, respectively. The dataset covers different regions (urban, residential, nature, and industrial), times (day and night), and weather conditions (sunny, rainy, and cloudy), presenting a significant challenge for evaluating generalization capability of the perception and localization methods (Li et al. 2022; Qiao et al. 2023). Only 6 surrounding cameras are used in the visual localization task. During the training and validation process, we add random pose deviations to real vehicle poses as initial vehicle poses. To fairly and fully compare our methods with existing localization algorithms on nuScenes dataset (Caesar et al. 2020), we follow three evaluation setups to determine the distributions of randomly generated longitudinal, lateral, and yaw deviations:

a) Default setup: the pose deviations are uniformly sampled from predefined ranges following (ZHANG et al. 2025), $\triangle x^* \sim \mathcal{U}(-2,2)$, $\triangle y^* \sim \mathcal{U}(-1,1)$, $\triangle \alpha^* \sim \mathcal{U}(-2,2)$, which is used as the default setup in the ablation studies.

b) RTMap setup: the pose deviations are sampled from predefined Gaussian distributions as in (Du et al. 2025), $\triangle x^* \sim \mathcal{N}(0, 1.5^2)$, $\triangle y^* \sim \mathcal{N}(0, 0.75^2)$, $\triangle \alpha^* \sim \mathcal{N}(0, 0.85^2)$.

c) LG-FA setups: the pose deviations are sampled from predefined Gaussian distributions following (Yong et al. 2026). Setup A uses $\triangle x^* \sim \mathcal{N}(0, 1^2)$, $\triangle y^* \sim \mathcal{N}(0, 1^2)$, $\triangle \alpha^* \sim \mathcal{N}(0, 2^2)$ whereas setup B increases the standard deviation by a factor of two. We do not change the parameter settings of the **RAVE** network for RTMap setup and LG-FA setups, in order to test its robustness against initial pose errors that exceed the **DEMA** sampling range. However, to achieve better performance, the **RAVE** model is trained for 35 epochs on these setups.

We measure the Mean Absolute Errors (MAE) of estimated 3 DoF poses (ZHANG et al. 2025), and Intersection-over-Union (IoU) scores of perceived static map elements (Li et al. 2022) for quantitative evaluation and comparison.

4.2. Comparison with the State-of-the-Arts

In order to prove the effectiveness and novelty of the proposed E2E localization network **RAVE**, we evaluate the localization accuracy in terms of the estimated 3 DoF poses on the nuScenes dataset (Caesar et al. 2020) and compare our method with existing rule-based visual localization methods (Yong et al. 2026) and E2E methods (ZHANG et al. 2025; Miao et al. 2025; Du et al. 2025). For fairness, we evaluate our proposed method under the same benchmarks used by the competing methods (ZHANG et al., 2025; Yong et al., 2026; Du et al., 2025), since these methods are tested under different evaluation setups and are not open-sourced. To fully demonstrate the superiority of **RAVE**, we additionally provide the performance of several methods evaluated under other benchmarks (Sarlin et al. 2023; Zhou et al. 2025; Zhao et al. 2024; Zhang et al. 2022). We also test the performance of the E2E localization network with different configurations, that is, the one using coarse localization **DEMA** only (Ours(C)), the full system without adaptive strategy (Ours(C+F)), and the adaptive localization system (Ours(A) or **RAVE**), as listed in Table 1.

Table 1. Comparison of localization accuracy with existing methods in terms of longitudinal, lateral and yaw errors on nuScenes dataset. "M." means monocular camera while "S." means surrounding cameras."*" means root mean square error.

Methods	E2E	Sensors	Map	Localization MAE↓		
				x (m)	y (m)	α (°)
Default Setup: $\triangle x^* \sim \mathcal{U}(-2,2), \triangle y^* \sim \mathcal{U}(-1,1), \triangle \alpha^* \sim \mathcal{U}(-2,2)$						
Init Pose				1.01	0.49	0.99
BEVLocator 2025	√	S.	HD	0.18	**0.08**	0.51
Miao et al. 2025	√	S.	HD	0.19	0.13	0.39
Ours (C)	√	S.	HD	0.15	0.13	0.38
Ours (C+F)	√	S.	HD	**0.09**	0.12	**0.35**
Ours (A)	√	S.	HD	0.11	0.13	**0.35**
RTMap Setup: $\triangle x^* \sim \mathcal{N}(0, 1.5^2), \triangle y^* \sim \mathcal{N}(0, 0.75^2), \triangle \alpha^* \sim \mathcal{N}(0, 0.85^2)$						
Init Pose				1.205	0.598	0.684
RTMap (E2E) 2025	√	S.	HD	0.589	0.142	0.521
RTMap (opt) 2025		S.	HD	0.586	**0.121**	0.368
Ours (C)	√	S.	HD	0.261	0.180	0.352
Ours (C+F)	√	S.	HD	**0.119**	0.145	**0.335**
LG-FA Setup A: $\triangle x^* \sim \mathcal{N}(0, 1^2), \triangle y^* \sim \mathcal{N}(0, 1^2), \triangle \alpha^* \sim \mathcal{N}(0, 2^2)$						
Init Pose				1.24	1.58	
LG-FA (ICP) 2026		S.	HD	1.51	1.22	
LG-FA (NDT) 2026		S.	HD	0.62	0.96	
LG-FA 2026		S.	HD	0.33	**0.24**	
Ours (C)	√	S.	HD	0.38	0.68	
Ours (C+F)	√	S.	HD	**0.27**	0.46	
LG-FA Setup B: $\triangle x^* \sim \mathcal{N}(0, 2^2), \triangle y^* \sim \mathcal{N}(0, 2^2), \triangle \alpha^* \sim \mathcal{N}(0, 4^2)$						
Init Pose				2.52	3.22	
LG-FA (ICP) 2026		S.	HD	3.27	1.92	
LG-FA (NDT) 2026		S.	HD	1.29	1.34	
LG-FA 2026		S.	HD	0.72	0.85	
Ours (C)	√	S.	HD	1.24	1.89	
Ours (C+F)	√	S.	HD	**0.41**	**0.67**	
Other Baselines using Different Evaluation Setups						
OrienterNet 2023	√	M.	SD	14.79	46.04	
SegLocNet 2025	√	S.	SD	8.15	19.68	
SegLocNet 2025	√	S.	HD	5.30	10.11	
Zhao et al. (ICP) 2025		S.	HD	0.39*	-	
Zhao et al. 2025		S.	HD	0.34*	-	
Zhang et al. 2022		M.	HD	0.22	0.10	

The quantitative experimental results show that our proposed method **RAVE** can achieve decimeter-level accuracy on the challenging nuScenes dataset (Caesar et al. 2020). Furthermore, the fine localization module **POET** can consistently refine the accuracy of estimated poses under all the setups, showing the effectiveness of the proposed coarse-to-fine localization framework.

Compared to LG-FA (Yong et al. 2026) that utilizes learning-based BEV perception but estimates pose by rule-based solvers, our work localizes vehicles in an E2E scheme through deep neural networks including both BEV perception and differentiable pose estimation, so the parameters of the pose estimation network can be automatically tuned throughout the E2E training stage, leading to comparable and even better robustness and localization accuracy. This advantage becomes more significant when the initial pose error is large (as shown in LG-FA setup B). It is noteworthy that we do not adjust the hyperparameters of the **RAVE** network according to the experimental setup. Ours(C) exhibits suboptimal performance in the LG-FA setup B, as the initial poses in many cases lie beyond its solvable range. Nevertheless, after incorporating the fine localization module, the pose estimation accuracy is significantly improved, substantially outperforming the competing methods. This indicates the advantages and potentials of the full E2E localization methods that after training on large volumes of real-world data, the network can generalize well on varying realistic driving scenarios without scene-specific parameter and rule tuning by experienced engineers.

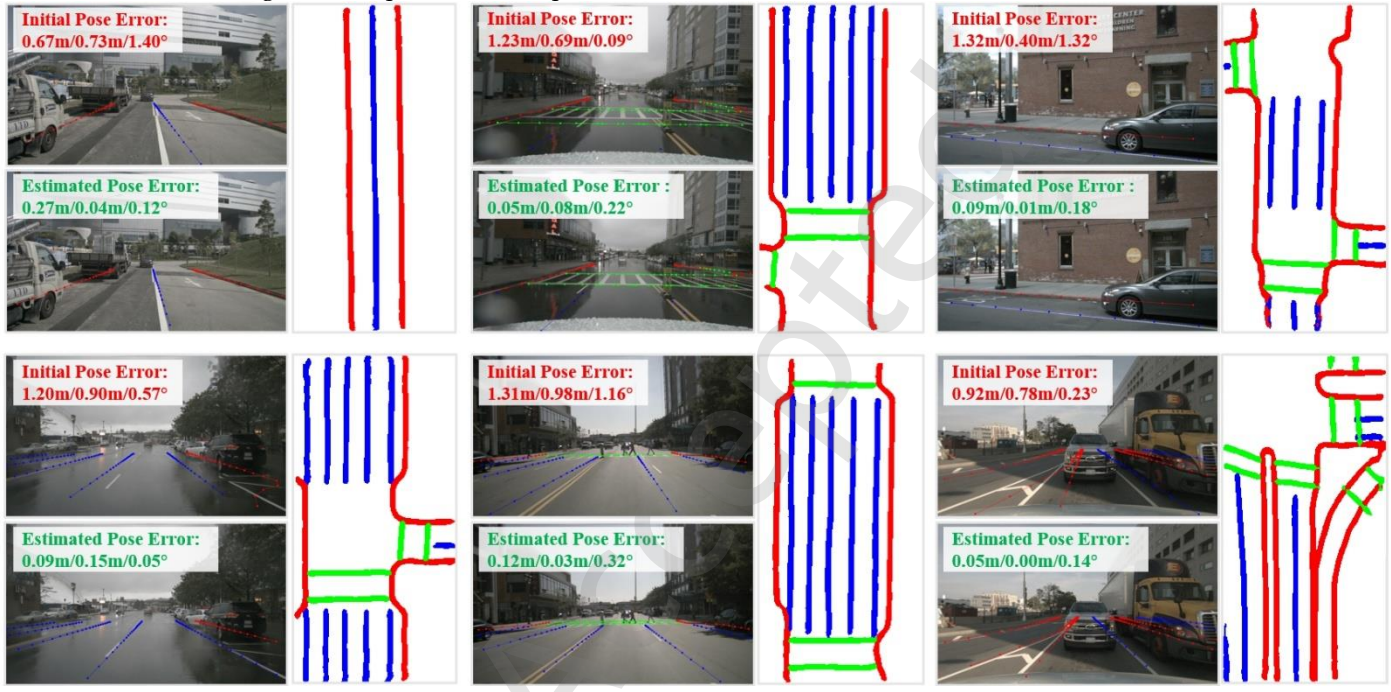

Fig. 7. Visualization cases of E2E localization on the nuScenes (Caesar et al. 2020). HD map elements are projected onto images viz initial pose (upper left) and estimated pose (bottom left). The perceived map elements are shown in BEV space (right). Road boundaries, lane dividers, and pedestrian crossings are drawn in red, blue, and green, respectively.

Table 2. Robustness towards various visual conditions. The improved performance by **POET** is highlighted by **BOLD**.

MAE ↓	Environmental Condition					
	Day	Night	Sunny	Cloudy	Rainy	All
Ours (C)						
x (m)	0.146	0.131	0.138	0.147	0.169	0.145
y (m)	0.131	0.142	0.134	0.124	0.134	0.132
α (°)	0.376	0.388	0.392	0.351	0.370	0.378
Ours (C+F)						
x (m)	**0.094**	**0.081**	**0.087**	**0.095**	**0.116**	**0.093**
y (m)	**0.121**	**0.133**	**0.124**	**0.114**	**0.125**	**0.122**
α (°)	**0.345**	**0.369**	**0.359**	**0.318**	**0.348**	**0.348**

Compared to fully E2E localization algorithms (ZHANG et al. 2025; Miao et al. 2025; Du et al. 2025), our work also demonstrates significantly improved localization accuracy. While Miao et al. utilized **DEMA** solver for efficient pose estimation (Miao et al. 2025), our work in this paper substantially extends it by introducing novel map prior-enhanced perception to improve the robustness of perception, and designing hierarchical localization modules to improve the accuracy of localization, especially in terms of longitudinal pose estimation. Both BEVLocator (ZHANG et al. 2025) and RTMap (Du et al. 2025) utilize full Transformer-based pose regressor and achieve superior lateral pose estimation accuracy to ours, but their performance on longitudinal position and yaw angle estimation is limited. Moreover, the whole black-box framework in BEVLocator (ZHANG et al. 2025) and RTMap (Du et al. 2025) lacks reliability and interpretability, which brings uncertainty to AD tasks. Our proposed **RAVE** algorithm first estimates coarse pose by interpretable **DEMA** solver and then refines the estimation by Transformer-based **POET** module, which achieves localization accuracy, efficiency, and interpretability.

We provide some qualitative visualization cases in Fig. 7. Our method can jointly perceive the static driving environment and perform localization under varying environmental conditions through a unified E2E network. By applying our method, vehicle pose can be accurately estimated, enabling precise alignment between projected map elements and realistic landmarks, even with large initial pose errors.

We also test the robustness of the proposed method towards various environmental conditions, as shown in Table 2. Taking advantage of the E2E framework, the network achieves stable pose estimation with high precision under varying conditions. In all of the evaluated conditions, the fine localization stage in the proposed hierarchical localization framework, *i.e.*, **POET** module, can refine

the estimated pose and improve the localization accuracy.

4.3. Ablation Analysis

4.3.1 Ablation Analysis about the FLORA Module

To analyze the proposed map prior-enhanced BEV perception method, **FLORA**, we evaluate its performance and determine its optimal configuration through quantitative experiments.

Effectiveness of the FLORA module: In this paper, BeMapNet (Qiao et al. 2023) is selected as our BEV perception backbone. We compare this baseline with some map prior-based perception enhancement works (Jiang et al. 2024; Zeng et al. 2024; Xiong et al. 2023; Jia et al. 2024) and our proposed **FLORA** method. Since BEV perception enhanced by rasterized HD map lacks existing methods, we also build some compared methods using well-known fusion techniques:

1) BeMapNet-ADD: $\mathbf{F}_{\mathcal{J}}^{e} = \mathbf{F}_{\mathcal{J}} + \mathbf{F}_{\mathcal{M}}^{R}$;
2) BeMapNet-CON: $\mathbf{F}_{\mathcal{J}}^{e} = \mathrm{conv}([\mathbf{F}_{\mathcal{J}}, \mathbf{F}_{\mathcal{M}}^{R}])$;
3) BeMapNet-ATT: $\mathbf{F}_{\mathcal{J}}^{e} = \mathrm{crossattn}\ (\mathbf{F}_{\mathcal{J}}, \mathbf{F}_{\mathcal{M}}^{R})$;
4) BeMapNet-GRU: $\mathbf{F}_{\mathcal{J}}^{e} = \mathrm{GRU}(\mathbf{F}_{\mathcal{J}}, \mathbf{F}_{\mathcal{M}}^{R})$;
5) BeMapNet-DGRU: $\mathbf{F}_{\mathcal{J}}^{e} = \mathrm{DeformableGRU}(\mathbf{F}_{\mathcal{J}}, \mathbf{F}_{\mathcal{M}}^{R})$;
6) BeMapNet-FLOW acts as **FLORA**, but it does not perform low-rank flow feature projection.

Table 3. Comparison of online BEV perception performance with existing methods. "S" means SD map prior, "H" means HD map prior, and "N" means neural map prior.

Methods	BEV perception performance (IoU) ↑				
	Bg.(%)	Div.(%)	Ped.(%)	Boud.(%)	Avg.(%)
Vision-only solution					
HDMapNet	-	40.6	18.7	39.5	-
BeMapNet	88.5	46.5	27.0	46.0	52.0
Map prior enhanced solution					
P-MapNet (+S+H)	-	44.3	23.3	43.8	-
PriorDrive (+H)	-	42.3	21.2	44.9	-
NMP (+N)	-	55.0	34.1	56.5	-
DiffMap (+H)	-	42.1	23.9	42.2	-
BeMapNet-ADD (+H)	90.5	57.9	27.5	52.7	57.1
BeMapNet-CON (+H)	90.3	56.9	26.4	51.8	56.4
BeMapNet-ATT (+H)	88.0	46.8	34.5	48.3	54.4
BeMapNet-GRU (+H)	90.4	57.5	28.4	53.8	57.6
BeMapNet-DGRU(+H)	88.1	47.0	57.0	48.0	60.0
BeMapNet-FLOW(+H)	93.1	65.8	65.2	66.1	72.5
Ours (+H)					
BeMapNet-FLORA (4)	91.5	59.3	63.6	60.1	68.6
BeMapNet-FLORA (8)	91.8	60.1	64.8	61.1	69.4
BeMapNet-FLORA(12)	92.3	62.5	66.2	62.8	71.0
BeMapNet-FLORA(16)	**93.7**	**67.9**	**68.6**	**68.4**	**74.6**
BeMapNet-FLORA(20)	91.3	58.3	63.9	59.5	68.2
BeMapNet-FLORA(24)	89.4	53.8	41.7	51.0	59.0

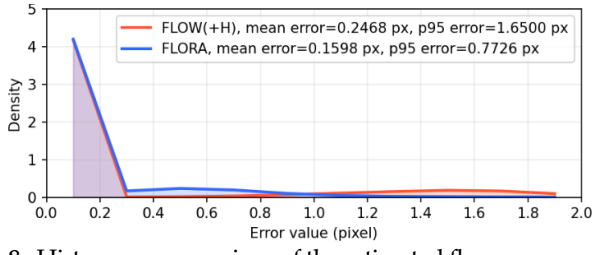

Fig. 8. Histogram comparison of the estimated flow error on map element areas between our proposed **FLORA** module and the vanilla FLOW module. For visualization clarity, we truncate the maximum error at two pixels. In fact, the proportion of errors exceeding two pixels is very small.

The experimental results are listed in Table 3. From the table, we can see that the BEV perception performance is consistently improved by the map prior. P-MapNet (Jiang et al. 2024) and DiffMap (Jia et al. 2024) learn HD map prior in their training stage, resulting in limited perception enhancement. PriorDrive (Zeng et al. 2024) fuses the HD map prior online, but uses only the road boundaries. In localization tasks with available HD maps, these methods cannot fully leverage the HD map prior to enhance online BEV perception. However, without precise vehicle pose, the retrieved HD map is not accurately aligned to online perception. Directly fusing HD map prior leads to unpleasant performance, as shown by BeMapNet-ADD, BeMapNet-CON, BeMapNet-ATT, and BeMapNet-GRU in Table 3. If the spatial displacement between retrieved HD map and online perception is solved, the perception enhancement will be significant, like BeMapNet-DGRU and BeMapNet-FLOW.

Experimental results demonstrate the proposed **FLORA** module achieves the best performance in BEV perception. With the low-rank flow feature projection, the dominant flow feature can be retained, while noise will be reduced, enabling more accurate alignment between HD map prior and online perception, leading to better enhancement than vanilla flow. As visualized in Fig. 8, the proposed **FLORA** module can effectively reduce the estimated flow errors. The experiment shows the **FLORA** module improves perception performance in terms of IoU scores from 52% to 74.6%.

We also evaluate the perception enhancement by **FLORA** module under varying environmental conditions, as shown in Table 4 and Fig. 9. The perception performance of original BeMapNet (Qiao et al. 2023) in terms of difficult map elements (such as pedestrian crossing) decreases dramatically under poor lighting (night), cloudy, and rainy conditions. With **FLORA** module, the HD map prior effectively enhances online perception so that the system can perceive stably and accurately under different conditions with a consistent improvement of more than 20% in terms of IoU scores.

Table 4. The improved robustness toward various visual conditions using the **FLORA** module.

Conditions	BEV perception performance (IoU) ↑				
	Bg. (%)	Div. (%)	Ped. (%)	Boud. (%)	Avg. (%)
BeMapNet					
Day	88.1	46.5	27.3	46.0	52.0
Night	91.7	46.9	12.3	45.2	49.0
Sunny	87.9	45.3	30.1	46.9	52.6
Cloudy	88.8	48.6	24.2	47.9	52.4
Rainy	87.4	46.7	22.9	39.7	49.2
BeMapNet-FLORA (16)					
Day	93.5	67.9	68.6	68.4	74.6
Night	95.3	67.9	70.6	68.5	75.6
Sunny	93.3	67.2	69.2	68.3	74.5
Cloudy	93.8	69.4	68.0	69.0	75..0
Rainy	93.3	67.7	67.3	67.4	73.9

Effectiveness of Flow Feature Rank: In the proposed **FLORA** module, the high-dimensional flow features are projected into the subspace built by low-rank flow features bases to suppress noise in flow estimation. To determine the optimal configuration of the **FLORA** module, we regulate the rank of flow feature, K, and test

the perception performance of the variants, as in Table 3. According to the results, as flow feature rank increases, the perception performance initially improves and then declines, showing that a proper rank of flow feature is necessary for flow estimation. Based on the experimental results, the **FLORA** module with $K = 16$ achieves the best perception performance, which is chosen as the default configuration in this paper.

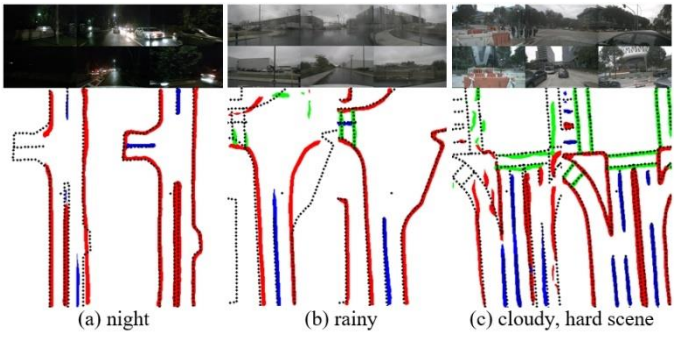

Fig. 9. Visualization cases of the enhanced BEV perception on the nuScenes (Caesar et al. 2020). The upper row shows surrounding images. The perceived map elements by original BeMapNet (Qiao et al. 2023) (bottom row: left) and BeMapNet-FLORA (bottom row: right) are shown in BEV space. Road boundaries, lane dividers, and pedestrian crossings are drawn in red, blue, and green, respectively. Ground-truth HD map points are visualized using black circles.

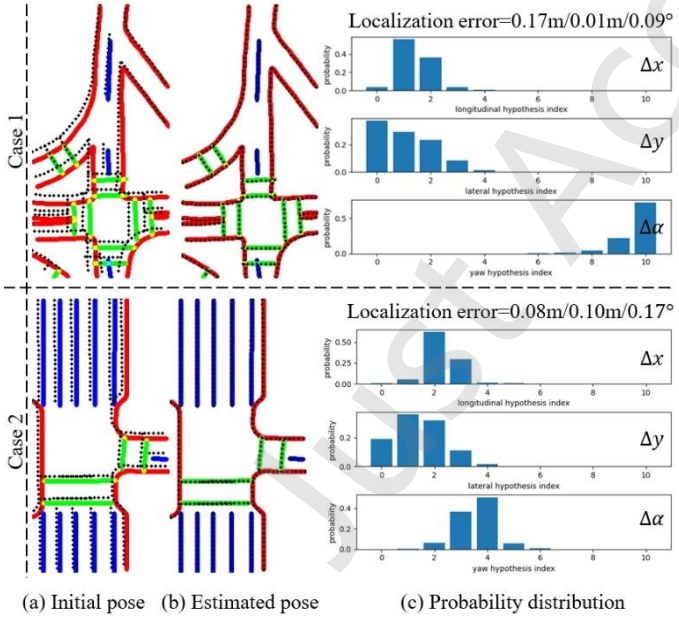

Fig. 10. Some examples of the **DEMA** solver. (a) illustrates the map projection using initial pose, (b) displays the map projection using estimated pose, (c) presents the estimated probability distribution of the 3 DoF poses. In (a) and (b), projected HD map points are drawn in black circles.

4.3.2 Ablation Analysis about the DEMA Module

To analyze the proposed **DEMA** solver and conclude its novelty, we compare its performance with traditional **FUMA** solver and then evaluate the performance of variants with different inputs, different sizes of input features, and different sampling steps. To intuitively visualize the estimated probabilities of 3 DoF poses by **DEMA** solver, we provide some examples in Fig. 10. The decoupled pose hypothesis with the highest probability is selected as the final estimation of **DEMA** solver. For rapid evaluation, the **FUMA** solver (F) and **DEMA** solver (D) evaluated in this part are trained for only 5 epochs in the second E2E training stage without adding the **POET** module.

FUMA versus DEMA: First, we compare the localization performance and computational efficiency between traditional **FUMA** solver and the proposed **DEMA** solver. The results are listed as Table 5 (F.a)-(D.c) and (F.d)-(D.g). By decoupling the effects on BEV features by each DoF of poses, the **DEMA** solver can individually solve 3 DoF poses by neural matching in a divide-and-conquer manner, reducing the number of pose hypotheses from $N_x \times N_y \times N_\alpha$ to $N_x + N_y + N_\alpha$. In the experiments, the **DEMA** solver can achieve a comparable or even better localization accuracy than **FUMA** solver while significantly reducing around 96% inference memories required by the pose solver, indicating that the proposed **DEMA** solver is an efficient and effective alternative to traditional **FUMA** solver.

Table 5. The localization performance using various pose solver and different inputs. The inference memory required by the pose solver is measured by subtracting the GRU memory of BEV perception from that of the entire network.

Solver	FLORA	Inputs	D_c	Size	Step (m/m/°)	Localization MAE↓ x (cm)	y (cm)	α (°)	Memory (MiB)
(F.a)		Features	16	100×50	0.4/0.2/0.4	17.3	16.4	0.411	64.31
(F.b)		Features	16	100×50	0.2/0.1/0.2	14.5	15.7	0.397	14888.95
(F.c)		Features	16	100×50	0.8/0.4/0.8	25.2	18.2	0.453	8.73
(F.d)	✓	Features	16	100×50	0.4/0.2/0.4	15.9	14.9	0.427	66.15
(D.a)		Results	4	100×50	0.4/0.2/0.4	23.1	18.6	0.533	1.82
(D.b)	✓	Results	4	100×50	0.4/0.2/0.4	16.7	14.8	0.496	1.85
(D.c)		Features	16	100×50	0.4/0.2/0.4	16.1	16.7	0.495	2.53
(D.d)		Features	16	100×50	0.2/0.1/0.2	13.1	16.3	0.502	3.67
(D.e)		Features	16	100×50	0.8/0.4/0.8	24.9	19.4	0.517	1.96
(D.f)		Features	16	400×200	0.4/0.2/0.4	16.2	15.8	0.462	22.84
(D.g)	✓	Features	16	100×50	0.4/0.2/0.4	16.0	14.2	0.424	2.14
(D.h)		Features	64	100×50	0.4/0.2/0.4	15.5	15.4	0.490	10.69
(D.i)	✓	Features	64	100×50	0.4/0.2/0.4	15.0	13.6	0.419	11.00
(D.j)		Features	128	100×50	0.4/0.2/0.4	15.2	14.6	0.471	36.15
(D.k)	✓	Features	128	100×50	0.4/0.2/0.4	15.1	13.6	0.422	35.36

Perception Features versus Perception Results: We analyze the desirable inputs of the **DEMA** solver by feeding BEV features (Table 5 (D.c), (D.g), (D.h), (D.i), (D.j), and (D.k)) or perception results (Table 5 (D.a) and (D.b)). By comparing the performance of (D.a) and (D.b), we can see that the dimension of perception results (perceived static map mask) is as low as 4, the perception results are not stable without map prior enhancement, so the pose solver using perception results works much worse than the one using BEV features. If we increase the dimension D_c of BEV features fed into the pose solver as (D.h) and (D.j), the localization accuracy consistently improves, indicating that higher-dimensional features are more stable and accurate, but the computational consumption will naturally increase. When incorporating the **FLORA** module into BEV perception, the quality of visual perception is significantly enhanced, so the localization performance of pose solvers using both BEV features and perception results is improved in all cases, such as (D.a)-(D.b), (D.c)-(D.g), (D.h)-(D.i), and (D.j)-(D.k). This

concludes that the necessity of BEV perception enhancement in this work lies in the fact that the E2E localization requires accurate and stable BEV perception as inputs. By considering both the localization accuracy and computational efficiency, we finally select the **DEMA** solver using enhanced BEV feature with dimension $D_c = 16$ as the default configuration during coarse localization.

Effectiveness of Sampling Step: We also test the effects of regulating sampling step of both **FUMA** solver and **DEMA** solver, as shown in Table 5 (F.b), (F.c), (D.d), and (D.e). Refining sampling step increases the number of pose hypotheses in both solvers, especially in traditional **FUMA** solver. The number of pose hypotheses in **FUMA** solver is increased from $N_x \times N_y \times N_\alpha$ to $8 \times N_x \times N_y \times N_\alpha$, resulting in an unbearable increase in inference memory. The increased computational burden of **DEMA** solver is not as significant as that of the **FUMA** solver, since its pose hypotheses are only increased linearly. Although refining the sampling step improves the localization performance of both **FUMA** solver and **DEMA** solver, its improvement is not stable. For example, the estimation accuracy of (D.d) in yaw angle is slightly worse than that of (D.c). Therefore, we select $0.4\text{m}/0.2\text{m}/0.4°$ as the default sampling step in the **DEMA** solver for both accuracy and efficiency.

Effectiveness of BEV Size: Before feeding the BEV features into the **DEMA** solver, we compress the size of BEV features for efficiency. We test the performance of a variant using BEV features with large size (*e.g.*, $H_c \times W_c = 400 \times 200$), as shown in Table 5 (D.f). It seems that using upsampled BEV features slightly improves the localization accuracy. We believe that larger features in BEV space can provide more detailed information, which help the model align the perceived features with the map features. However, enlarging BEV features imposes a heavy computational burden for pose estimation. Therefore, we keep using compressed BEV features in the proposed E2E localization network.

4.3.3 Ablation Analysis about the POET Module

In this paper, we propose to utilize a Transformer-based pose regressor **POET** to refine the coarse estimation in this fine localization stage. We conduct detailed analysis of its performance.

Effectiveness of the POET module: We first validate the effectiveness of the proposed **POET** module, as shown in Table 6. In the experiments, we found that the **DEMA** solver converges rapidly yet fails to improve consistently with extended training, which stems from the limited resolution of BEV perception and rasterized map encoding. Using the **POET** module to refine the coarse estimation from the **DEMA** solver can effectively further improve the localization accuracy, especially in terms of longitudinal pose estimation, which is necessary for high-level AD systems. We attribute this to the utilized fine-grained map features and deep Transformer-based network structures. Although the **POET** module converges slower and requires more computational resources than the **DEMA** solver, it has better potential to achieve higher localization accuracy.

To better understand the role of the **POET** network architecture and vectorized map features in the fine localization module, we conduct more detailed ablation studies, as shown in Table 7. We first build three variants of pose regressors based on vectorized map features, namely, MLP, convolutional network (Conv.), and vanilla Transformer (Trans.). To ensure a fair comparison, the network parameters of the three variants are kept almost identical to those of **POET**, so as to demonstrate that the performance improvement does not come simply from introducing additional parameters. As the results show, with the same vectorized map features, all pose regressors can further refine the pose and achieve higher localization accuracy than Ours(C). However, **POET** significantly outperforms the other three variants, especially in longitudinal position optimization, demonstrating the superiority of the **POET** architecture. Furthermore, we also replace the vectorized map features with rasterized map features as input of the fine localization module. To process rasterized map features, we have to use a heavy rasterized encoder instead of the lightweight vectorized encoder, resulting in a substantial increase in model parameters. Moreover, it is clear that all pose regressors struggle to further refine the pose, with localization accuracy significantly lower than that of pose regressors using vectorized map features. This validates our hypothesis that vectorized map features provide higher accuracy than rasterized features, thus enabling further improvement of localization performance.

Table 6. Comparison of localization accuracy with or without refinement by the fine localization module. Only the computation time of hierarchical localization module is measured here.

Methods	t_c	Activation Ratio (%)	Localization MAE↓ x (cm)	y (cm)	α (°)	Time (ms)
Ours (C)	0	0.0	14.5	13.2	0.378	6.14
Ours (C+F)	1.0	100.0	9.3	12.2	0.348	19.39
Ours (A)	0.2	4.4	14.3	13.2	0.375	6.73
	0.3	30.2	13.0	13.0	0.367	10.15
	0.4	74.5	10.6	12.5	0.354	16.01
	0.5	97.5	9.4	12.3	0.348	19.06
	0.6	99.9	9.3	12.2	0.348	19.38

Table 7. Comparison of localization accuracy with different pose regressors using vectorized or rasterized map features.

Regressor	Map Feature	Model Size	Localization MAE↓ x (cm)	y (cm)	α (°)
POET	Vectorized	2.44 M	**9.3**	**12.2**	**0.348**
POET	Rasterized	4.18 M	14.6	13.5	0.371
MLP	Vectorized	2.37 M	12.4	13.1	0.360
MLP	Rasterized	4.11 M	15.4	13.8	0.380
Conv.	Vectorized	3.78 M	14.0	12.5	0.353
Conv	Rasterized	5.53 M	15.0	13.9	0.393
Trans.	Vectorized	2.29 M	12.1	12.3	0.351
Trans.	Rasterized	4.03 M	14.4	13.4	0.375

Fig. 11 demonstrates the **POET** module refines the coarse estimation. Due to the limited resolution of BEV features, the localization accuracy of **DEMA** solver is near the resolution of BEV perception ($0.15\text{m}/\text{pixel}$), which supports spatial alignment between two BEV features. After refinement by the **POET** module, the localization accuracy of the system is further improved, achieving a centimeter-level high-precision pose estimation.

4.3.4 Ablation Analysis about the Adaptive Coarse-to-Fine Strategy

We analyze the effectiveness of the adaptive strategy in the proposed hierarchical localization system. By regulating the threshold t_c, we can control the activation frequency of the **POET**

module, as shown in Table 6 and Fig. 12. As the threshold increases, more and more coarse estimation results are regarded as unreliable, leading to an increase in the frequency of using the **POET** module to optimize the results. Thus, the localization accuracy is consistently improved by the added fine localization stage, but the computational efficiency is sacrificed. Finally, considering the balance between localization accuracy and efficiency, we select $t_c = 0.4$ as the default threshold in the adaptive strategy.

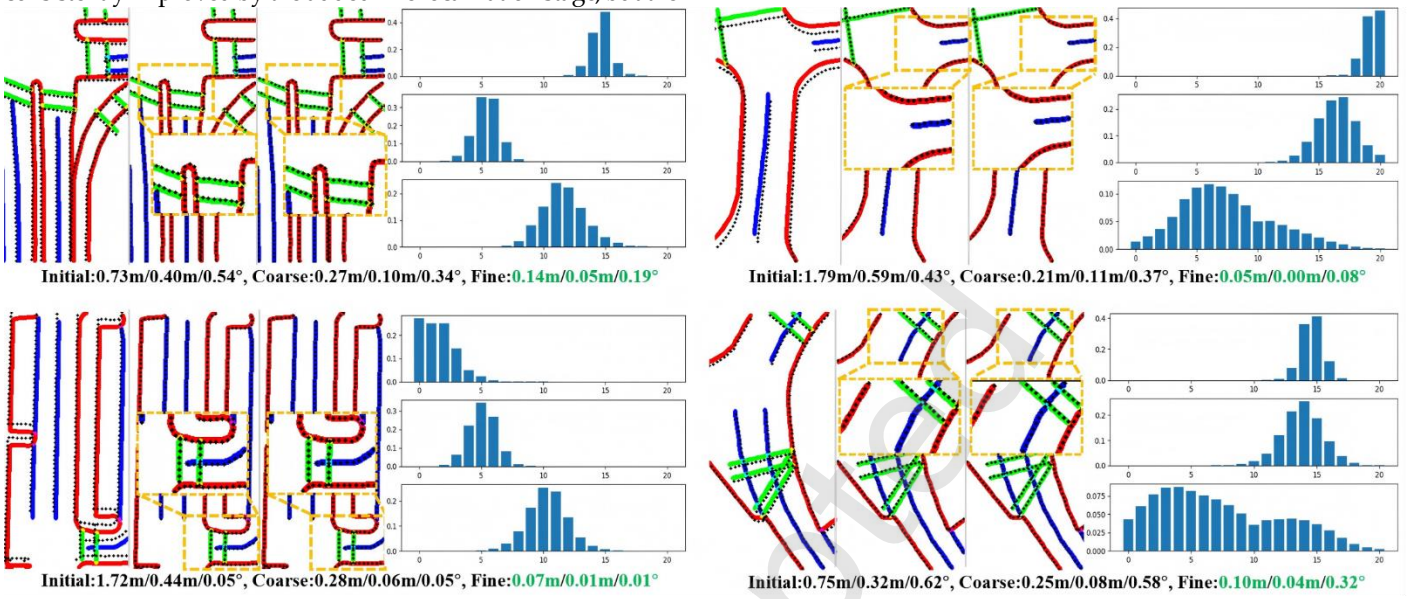

Fig. 11. Examples demonstrating the **POET** module refines the estimation of the **DEMA** solver. In each case, the map projection using initial pose, coarse pose estimated by **DEMA** solver, and fine pose estimated by **POET** module are shown in black circles from left to right; ground-truth HD map masks are drawn in blue, green, and red lines, corresponding to lane dividers, pedestrian crossings, and road boundaries, respectively. Estimated probability distributions from **DEMA** solver are also provided.

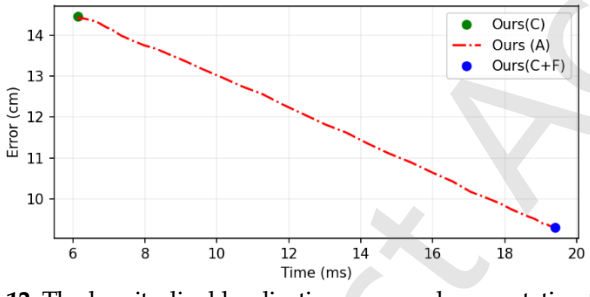

Fig. 12. The longitudinal localization error and computation time curve across different thresholds.

Table 8. Localization accuracy using missing map elements and poor camera calibration.

Map missing ratio	Coarse Localization MAE↓			Full Localization MAE↓		
	x (cm)	y (cm)	α (°)	x (cm)	y (cm)	α (°)
0.0	14.5	13.2	0.378	9.3	12.2	0.348
0.1	15.0	13.7	0.394	9.9	12.7	0.363
0.2	16.4	15.9	0.454	11.7	15.1	0.421
0.3	18.3	17.6	0.498	13.8	17.0	0.465
0.3 (trained)	16.7	15.0	0.432	12.6	14.0	0.402
Calib noise ratio	**Coarse Localization MAE↓**			**Full Localization MAE↓**		
	x (cm)	y (cm)	α (°)	x (cm)	y (cm)	α (°)
0.0	14.5	13.2	0.378	9.3	12.2	0.348
0.1	14.6	15.9	0.447	9.4	15.2	0.414
0.1 (trained)	14.4	13.5	0.389	9.5	12.2	0.358
BEVLocator 2025	-	-	-	18.0	8.0	0.510
Miao et al. 2025	19.0	13.0	0.39	-	-	-

4.3.5 Robustness Evaluation against Various Interferences

To test the robustness of the proposed **RAVE** method under incorrect map information, we simulate evaluation conditions with randomly missing map elements. We randomly masked out a fixed ratio of map elements during localization. We first tested the **RAVE** model without retraining, as shown in Table 8. Since the model uses complete map information during training, it strives to match all perceived map elements with the map information. Therefore, as the missing ratio of map elements increases, the discrepancy between the actual perception results and the map information grows, and the localization performance of the model gradually degrades. However, even when the missing ratio of map elements reaches as high as 30%, our **RAVE** model still achieves localization accuracy comparable to or even exceeding that of existing methods (ZHANG et al. 2025; Miao et al. 2025), and the fine localization module consistently further refines the pose accuracy. When we also mask out some map elements during training, allowing the model to learn to selectively use perception results for matching and localization against the map information, it can be seen that the localization accuracy of the **RAVE** method improves effectively under map missing conditions, and its robustness against incorrect map information is enhanced.

We also evaluated the **RAVE** model under inaccurate camera calibration. Specifically, we scale the intrinsic and the translation components of the extrinsic parameters by a factor of 1.1 to simulate calibration errors. As shown in Table 8, the performance of both coarse and fine localization modules degrades significantly under inaccurate calibration, particularly in lateral position and yaw angle estimation. Camera calibration directly affects BEV perception. While map priors can mitigate this interference to some degree, inaccurate BEV features still impair the spatial alignment and fusion of map priors. As a result, the visual observations for

localization degrade, reducing localization accuracy. However, when the model is retrained under such interference, we observe that the impact of calibration errors is substantially mitigated. This demonstrates that the end-to-end localization network can learn to correct perception errors caused by calibration inaccuracies, exhibiting good robustness.

Table 9. Multi-sensor fusion localization accuracy.

Method	Localization MAE↓		
	x (m)	y (m)	α (°)
IMU Dead Reckoning	35.504	35.857	0.416
GNSS+IMU	0.714	0.697	0.764
GNSS+IMU+Ours(C)	0.240	0.244	0.329
GNSS+IMU+Ours(C+F)	**0.226**	**0.235**	**0.306**

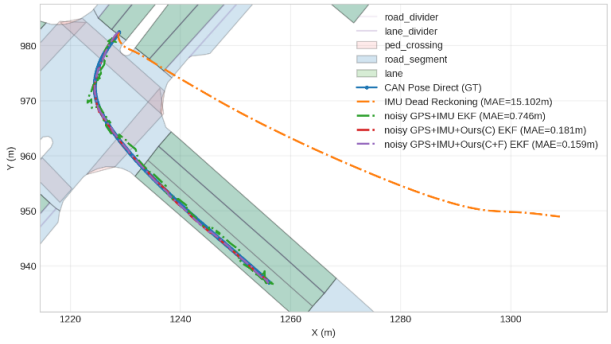

Fig. 13. The vehicle poses estimated by several localization system. We render the utilized HD map information through nuScenes API (Caesar et al. 2020) for visualization.

4.4. Multi-sensor Fusion Localization Applications

To evaluate the effectiveness of our method in a practical multi-sensor fusion localization system, we design an Extended Kalman Filter (EKF)-based GNSS+IMU fusion localization system. Since the nuScenes dataset (Caesar et al. 2020) does not provide raw GNSS signals, we follow common practices (ZHANG et al. 2025; Miao et al. 2025) by adding noise to the ground-truth keyframe poses at a low frequency (~2 Hz) to simulate noisy GNSS positioning signals, which are then fused with high-frequency (~100 Hz) IMU measurements as the baseline. Our proposed **RAVE** method is then applied to estimate the pose from the noisy GNSS signals, reducing their localization error before fusion. The results are shown in Table 9. It can be seen that our method significantly reduces GNSS localization errors, thus enabling accurate pose state updates within the EKF framework and constraining the gradually drifting IMU-integrated pose. Furthermore, the fine localization module **POET** further improves the localization accuracy, demonstrating the effectiveness of our coarse-to-fine localization framework. As illustrated in the visual examples in Fig. 13, although IMU integration provides continuous, high-frequency pose estimates, it inevitably suffers from accumulated drift over long time spans. Adding GNSS signal fusion reduces the drift, but raw GNSS signals have large errors, resulting in inaccurate and unsmooth trajectories. After pose estimation with our **RAVE** method, the localization error is significantly reduced, allowing the autonomous vehicle to estimate its pose over long time spans. Combined with high-frequency IMU pose prediction, the vehicle pose estimation frequency is also substantially increased, meeting the real-time pose estimation requirements of the vehicle. Moreover, under various traffic scenarios, even at complex intersections, our proposed **FLORA** module ensures high-quality BEV perception, so the **RAVE** method can consistently reduce localization errors and achieve high-precision pose estimation.

Table 10. Localization accuracy on self-collected dataset.

Method	Localization MAE↓		
	x (m)	y (m)	α (°)
Keyframe Pose Accuracy (~2 Hz)			
GNSS	1.91	4.49	0.96
GNSS+ Ours(C)	0.23	0.78	0.57
GNSS+ Ours(C+F)	**0.12**	**0.12**	**0.21**
Trajectory Accuracy (~200 Hz, ~17.1 km)			
GNSS+IMU	1.29	4.72	0.95
GNSS+ IMU+Ours(C)	0.26	0.79	0.54
GNSS+ IMU+Ours(C+F)	**0.14**	**0.15**	**0.25**

4.5. Real-world Vehicle Evaluation

In real-world applications, GNSS positioning errors are influenced by multiple factors and cannot be adequately characterized by a single Gaussian or uniform distribution. To demonstrate the practical value of the proposed **RAVE** method, we conduct a real-world vehicle test. A test vehicle is assembled, equipped with a GNSS, an IMU, 5 surrounding cameras, and a LiDAR sensor. We collect data over a distance of approximately 96.8 kilometers. In total, we obtain 12,400 keyframes, which are split into training and test sets at a ratio of roughly 5:1. The HD map used for localization is constructed from LiDAR data. During training, we still model the initial pose error as a Gaussian distribution. But, during evaluation, we retrieve map data and estimate vehicle poses based on real GNSS signals, thereby evaluating the localization performance of **RAVE** model under realistic GNSS disturbances.

The experimental results are listed in Table 10. We first measure the localization accuracy of keyframes at 2 Hz. It can be clearly observed that both the coarse and fine localization module in **RAVE** method reduce the localization error induced by real-world GNSS signals and the **POET** module consistently improves the performance. In addition, we integrate the **RAVE** method into the same EKF framework as described in Sec. 4.4, and evaluate the overall trajectory accuracy at a high frequency. Due to the long test distance, pure IMU dead reckoning method suffers from significant accumulated errors, whereas GNSS absolute poses provide effective constraints. With the proposed **RAVE** method mitigating GNSS positioning errors, the trajectory accuracy of the EKF-based fusion system is effectively enhanced, maintaining decimeter-level localization accuracy over a distance of up to 17 kilometers. The results demonstrate the effectiveness of the proposed **RAVE** method, confirming its practical validity and robustness in real-world applications.

5 Conclusions

In this paper, we address the visual-to-HD map localization problem by proposing an E2E hierarchical localization network **RAVE**, which ensures interpretability and computational efficiency while achieving decimeter-level localization accuracy. Surrounding images and HD map data are converted into BEV

features, rasterized and vectorized map features through neural networks, which overcomes the challenges of data association between cross-modal data. To improve the robustness of vision-based BEV perception, the proposed **FLORA** module incorporates the misaligned map prior information into the online perceived BEV features, which guarantees stable inputs for subsequent pose estimation. Then, a hierarchical localization module is proposed to solve vehicle poses, where the **DEMA** module efficiently estimates coarse pose using rasterized map features, and the **POET** module refines the results to high precision using vectorized map features. The proposed network is fully analyzed on the nuScenes dataset (Caesar et al. 2020) and is concluded to be able to achieve high-precision localization with a MAE of 0.106m, 0.125m, and 0.354° in longitudinal, lateral positions, and yaw angle. The interpretable **DEMA** module in coarse localization significantly reduces inference burden, and the **POET** module ensures the performance. By regulating the threshold of estimated probability in the adaptive strategy, the proposed network can achieve a good balance among localization accuracy, computational efficiency, and interpretability, which is essential for ego pose estimation for high-level safety-guarantee autonomous driving.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (52394264, 52472449, U22A20104, 52372414, 52402499), China Postdoctoral Science Foundation (2024M761636), and Independent Research Project of the State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University (ZZ-PY-20250408).

Author contributions

Jinyu Miao: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing - original draft.
Yifei He: Data curation, Software, Validation, Visualization, Writing - original draft.
Tuopu Wen: Funding acquisition, Investigation, Methodology, Supervision, Writing - review & editing.
Kun Jiang: Funding acquisition, Supervision, Writing - review & editing.
Kangan Qian, Zheng Fu, Yunlong Wang, and Zhihuang Zhang: Resources, Validation, Visualization.
Weitao Zhou: Formal analysis, Writing - review & editing.
Mengmeng Yang: Funding acquisition, Writing - review & editing.
Jin Huang, and Zhihua Zhong: Investigation, Resources, Writing - review & editing.
Diange Yang: Funding acquisition, Investigation, Supervision, Project administration, Writing - review & editing.

Replication and Data Sharing

The data and codes used in this study are available at https://doi.org/10.26599/ETSD.2026.9190077.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Jin Huang and Diange Yang are the Editorial Board Members of this journal.

References

Ali-bey, A., Chaib-draa, B., Giguere, P., 2024. Boq: A place is worth a bag of learnable queries, in: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17794–17803.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., 2018. Netvlad: Cnn architecture for weakly supervised place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 1437–1451.

Barnes, D., Weston, R., Posner, I., 2019. Masking by moving: Learning distraction-free radar odometry from pose information. arXiv preprint arXiv:1909.03752 .

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 11621–11631.

Camiletto, A.B., Bochicchio, A., Liniger, A., Dai, D., Gawel, A., 2024. U-bev: Height-aware bird's-eye-view segmentation and neural map-based relocalization, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5597–5604.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: Proceedings of European conference on computer vision (ECCV), Springer. pp. 213–229.

Cattaneo, D., Vaghi, M., Ballardini, A.L., Fontana, S., Sorrenti, D.G., Burgard, W., 2019. Cmrnet: Camera to lidar-map registration, in: Proceedings of IEEE intelligent transportation systems conference (ITSC), pp. 1283–1289.

Chen, Z., Xu, X., Wang, Y., Xiong, R., 2021. Deep phase correlation for end-to-end heterogeneous sensor measurements matching, in: Kober, J., Ramos, F., Tomlin, C. (Eds.), in: Proceedings of the 2020 Conference on Robot Learning (CoRL), PMLR. pp. 2359–2375.

Choi, M.J., Suhr, J.K., Choi, K., Jung, H.G., 2019. Low-cost precise vehicle localization using lane endpoints and road signs for highway situations. IEEE Access 7, 149846–149856.

Deng, L., Yang, M., Hu, B., Li, T., Li, H., Wang, C., 2019. Semantic segmentation-based lane-level localization using around view monitoring system. IEEE Sensors Journal 19, 10077–10086.

Ding, W., Qiao, L., Qiu, X., Zhang, C., 2023. Pivotnet: Vectorized pivot learning for end-to-end hd map construction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3672–3682.

Dong, H., Gu, W., Zhang, X., Xu, J., Ai, R., Lu, H., Kannala, J., Chen, X., 2024. Superfusion: Multilevel lidar-camera fusion for long-range hd map generation, in:

Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 9056–9062.

Du, Y., Yang, S., Wang, L., Hou, Z., Cai, C., Tan, Z., Chen, M., Huang, S.S., Li, Q., 2025. Rtmap: Real-time recursive mapping with change detection and localization, in: Proceedings of 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 28021-28030.

Fan, Y., Du, X., Luo, L., Shen, J., 2022. Fresco: Frequency-domain scan context for lidar-based place recognition with translation and rotation invariance, in: Proceedings of International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 576–583.

Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C., 2020. Vectornet: Encoding hd maps and agent dynamics from vectorized representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 11525–11533.

He, Y., Liang, S., Rui, X., Cai, C., Wan, G., 2024. Egovm: Achieving precise ego-localization using lightweight vectorized maps, in: Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 12248–12255.

Hosseinzadeh, M., Reid, I., 2024. Bevpose: Unveiling scene semantics through pose-guided multi-modal bev alignment, in: Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 14111–14118.

Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H., 2023. Planning-oriented autonomous driving, in: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17853–17862.

Jang, W., Hyun, J., An, J., Cho, M., Kim, E., 2021. A lane-level road marking map using a monocular camera. IEEE/CAA Journal of Automatica Sinica 9, 187–204.

Jia, P., Wen, T., Luo, Z., Yang, M., Jiang, K., Liu, Z., Tang, X., Lei, Z., Cui, L., Zhang, B., et al., 2024. Diffmap: Enhancing map segmentation with map prior using diffusion model. IEEE Robotics and Automation Letters.

Jiang, S., Wang, Y., Hu, Z., 2023. A data association method based on pose refinement and mismatch rejection for hd map localization, in: Proceedings of 2023 7th CAA International Conference on Vehicular Control and Intelligence (CVCI), pp. 1–6.

Jiang, Z., Zhu, Z., Li, P., Gao, H.a., Yuan, T., Shi, Y., Zhao, H., Zhao, H., 2024. P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. IEEE Robotics and Automation Letters.

Li, Q., Wang, Y., Wang, Y., Zhao, H., 2022. Hdmapnet: An online hd map construction and evaluation framework, in: Proceedings of 2022 International Conference on Robotics and Automation (ICRA), pp. 4628–4634.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J., 2024. Bevformer: learning bird's-eyeview representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C., 2023. Maptr: Structured modeling and learning for online vectorized hd map construction, in: Proceedings of International Conference on Learning Representations (ICLR).

Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X., 2024. Maptrv2: An end-to-end framework for online vectorized hd map construction. International Journal of Computer Vision, 1–23.

Liao, Z., Shi, J., Qi, X., Zhang, X., Wang, W., He, Y., Liu, X., Wei, R., 2020. Coarse-to-fine visual localization using semantic compact map, in: Proceedings of 2020 3rd International Conference on Control and Robots (ICCR), IEEE. pp. 30–37.

Liu, R., Wang, J., Zhang, B., 2020. High definition map for automated driving: Overview and analysis. The Journal of Navigation 73, 324–341.

Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H., 2023. Vectormapnet: End-to-end vectorized hd map learning, in: Proceedings of International Conference on Machine Learning (ICML), PMLR. pp. 22352–22369.

Miao, J., Jiang, K., Wang, Y., Wen, T., Xiao, Z., Fu, Z., Yang, M., Liu, M., Huang, J., Zhong, Z., et al., 2023. Poses as queries: End-to-end image-to-lidar map localization with transformers. IEEE Robotics and Automation Letters 9, 803–810.

Miao, J., Jiang, K., Wen, T., Wang, Y., Jia, P., Wijaya, B., Zhao, X., Cheng, Q., Xiao, Z., Huang, J., et al., 2024. A survey on monocular re-localization: From the perspective of scene map representation. IEEE Transactions on Intelligent Vehicles.

Miao, J., Wen, T., Luo, Z., Qian, K., Fu, Z., Wang, Y., Jiang, K., Yang, M., Huang, J., Zhong, Z., et al., 2025. Efficient end-to-end visual localization for autonomous driving with decoupled bev neural matching. In: Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10719-10726.

Pauls, J.H., Petek, K., Poggenhans, F., Stiller, C., 2020. Monocular localization in hd maps by combining semantic segmentation and distance transform, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4595–4601.

Pink, O., 2008. Visual map matching and localization using a global feature map, in: Proceedings of 2008 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), IEEE. pp. 1–7.

Qiao, L., Ding, W., Qiu, X., Zhang, C., 2023. End-to-end vectorized hd-map construction with piecewise bezier curve, in: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13218–13228.

Qu, X., Soheilian, B., Paparoditis, N., 2015. Vehicle localization using mono-camera and geo-referenced traffic signs, in: Proceedings of 2015 IEEE Intelligent

Vehicles Symposium (IV), IEEE. pp. 605–610.

Radenovic, F., Tolias, G., Chum, O., 2019. Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1655–1668.

Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From coarse to fine: Robust hierarchical localization at large scale, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 12716–12725.

Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., Balntas, V., 2023. Orienternet: Visual localization in 2d public maps with neural matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21632–21642.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Proceedings of 3rd International Conference on Learning Representations (ICLR).

Spangenberg, R., Goehring, D., Rojas, R., 2016. Pole-based localization for autonomous vehicles in urban scenarios, in: Proceedings of 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE. pp. 2161–2166.

Sun, R., Yang, L., Lingrand, D., Precioso, F., 2025. Mind the map! accounting for existing maps when estimating online hdmaps from sensors, in: Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1671–1681.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A., 2018. Inloc: Indoor visual localization with dense matching and view synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7199–7209.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R., 2020. Fourier features let networks learn high frequency functions in low dimensional domains, in: Proceedings of Advances in neural information processing systems (NeurIPS) 33, 7537–7547.

Tang, X., Jiang, K., Yang, M., Liu, Z., Jia, P., Wijaya, B., Wen, T., Cui, L., Yang, D., 2024. High-definition maps construction based on visual sensor: A comprehensive survey. IEEE Transactions on Intelligent Vehicles 9, 5973–5994.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), in: Proceedings of European conference on computer vision (ECCV), pp. 402–419.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), p. 6000–6010.

Vivacqua, R.P.D., Bertozzi, M., Cerri, P., Martins, F.N., Vassallo, R.F., 2017. Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving. IEEE Transactions on Intelligent Transportation Systems 19, 582–597.

Wan, J., Zhang, X., Dong, S., Zhang, Y., Yang, Y., Wu, R., Jiang, Y., Li, J., Lin, J., Yang, M., 2024. Monocular localization with semantics map for autonomous vehicles, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 14146–14152.

Wang, H., Xue, C., Zhou, Y., Wen, F., Zhang, H., 2021. Visual semantic localization based on hd map for autonomous vehicles in urban scenarios, in: Proceedings of 2021 IEEE International conference on robotics and automation (ICRA), IEEE. pp. 11255–11261.

Wen, T., Jiang, K., Wijaya, B., Li, H., Yang, M., Yang, D., 2022. Tm3loc: Tightly-coupled monocular map matching for high precision vehicle localization. IEEE Transactions on Intelligent Transportation Systems 23, 20268–20281.

Weston, R., Gadd, M., De Martini, D., Newman, P., Posner, I., 2022. Fast-mbym: Leveraging translational invariance of the fourier transform for efficient and accurate radar odometry, in: Proceedings of 2022 International Conference on Robotics and Automation (ICRA), IEEE. pp. 2186–2192.

Wu, H., Zhang, Z., Lin, S., Mu, X., Zhao, Q., Yang, M., Qin, T., 2024. Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps, in: Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 13198–13205.

Xiao, Z., Jiang, K., Xie, S., Wen, T., Yu, C., Yang, D., 2018. Monocular vehicle self-localization method based on compact semantic map, in: Proceedings of 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 3083–3090.

Xiao, Z., Yang, D., Wen, T., Jiang, K., Yan, R., 2020. Monocular localization with vector hd map (mlvhm): A low-cost method for commercial ivs. Sensors 20, 1870.

Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H., 2023. Neural map prior for autonomous driving, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17535–17544.

Ye, N., Wang, C., Fan, H., Liu, S., 2021. Motion basis learning for unsupervised deep homography estimation with subspace projection, in: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13097–13105.

Yuan, T., Liu, Y., Wang, Y., Wang, Y., Zhao, H., 2024. Streammapnet: Streaming mapping network for vectorized online hd map construction, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7356–7365.

Yue, H., Miao, J., Chen, W., Wang, W., Guo, F., Li, Z., 2022. Automatic vocabulary and graph verification for accurate loop closure detection. Journal of Field Robotics 39, 1069–1084.

Zeng, S., Chang, X., Liu, X., Pan, Z., Wei, X., 2024. Driving with prior maps: Unified vector prior encoding for autonomous vehicle mapping. arXiv preprint arXiv:2409.05352.

ZHANG, Z., XU, M., ZHOU, W., PENG, T., LI, L., POSLAD, S., 2025. Bev-locator: an end-to-end visual semantic localization network using multi-view images. SCIENCE CHINA Information Sciences 68, 122106–.

Zhang, Z., Zhao, J., Huang, C., Li, L., 2022. Learning visual semantic map-matching for loosely multi-sensor fusion localization of autonomous vehicles. IEEE Transactions on Intelligent Vehicles 8, 358–367.

Zhao, L., Liu, Z., Yin, Q., Yang, L., Guo, M., 2024. Towards robust visual localization using multiview images and hd vector map, in: Proceedings of 2024 IEEE International Conference on Image Processing (ICIP), pp. 814–820.

Zhou, Z., Qi, Z., Cheng, L., Xiong, G., 2025. Seglocnet: Multimodal localization network for autonomous driving via bird's-eye-view segmentation. arXiv preprint arXiv:2502.20077.

Wijaya, B., Yang, M., Jiang, K., Zhang, W., Yang, D., 2024. Exploring the application of blockchain technology in crowdsource autonomous driving map updating. Communications in Transportation Research 4, 100140.

Tang, X., Yang, M., Wen, T., Jia, P., Cui, L., Luo, M., Sheng, K., Zhang, B., Jiang, K., Yang, D., 2025. PriorFusion: Unified integration of priors for robust road perception in autonomous driving. Communications in Transportation Research 5, 100229.

Deng, N., Zhou, W., He, Y., Cheng, Q., Zhang, B., Wen, J., Liu, C., He, J., Sha, X., Qian, Z., Jiang, K., Yang, M., Cao, Z., Yang, D., 2026. Dynamically local-enhancement planner for large-scale autonomous driving. Communications in Transportation Research.

Yong, J., Qu, D., Chen, Q., Oguchi, K., Fukushima, S., 2026. Localization-Guided Foreground Augmentation in Autonomous Driving. arXiv preprint arXiv: 2604.18940.

Author biography

Jinyu Miao received his B.S. degree in Automation and his M.S. degree in Control Science and Engineering from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2019 and 2022, respectively. He is now currently working toward his Ph.D. degree at the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include localization, perception, and decision-making for autonomous driving.

Yifei He received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2024. He is now currently working toward his M.S. degree in Mechanical Engineering at the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include autonomous driving.

Tuopu Wen received the B.S. degree from Electronic Engineering, Tsinghua University, Beijing, China in 2018 and his Ph.D. degree at the School of Vehicle and Mobility of Tsinghua University in 2023. He is now working as a postdoctoral researcher in the School of Vehicle and Mobility of Tsinghua University. His research interests include computer vision, high-definition map, and high-precision localization for autonomous driving.

Kun Jiang received the B.S. degree in mechanical and automation engineering from Shanghai Jiao Tong University, Shanghai, China in 2011. He received the Master degree in mechatronics system and the Ph.D. degree in information and systems technologies from University of Technology of Compiegne (UTC), Compiegne, France, in 2013 and 2016, respectively. He is currently an Associate Research Professor at Tsinghua University, Beijing, China. His research interests include autonomous vehicles, high precision map, and sensor fusion.

Kangan Qian received his B.S. degree in Mechanical Design, Manufacturing and Automation from Central South University, Changsha, China, in 2023. He is now currently working toward his M.S. degree in Mechanical Engineering at the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include multimodal perception and reasoning.

Zheng Fu received the M.S. degree in pattern recognition and intelligent systems from the Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2019, and the Ph.D. degree from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2025. Her research focuses on the intersection of autonomous driving and embodied intelligence, dedicated to building multimodal world models that integrate vision, language, and tactile modalities to support environment understanding and intelligent decision-making in complex dynamic scenarios.

Yunlong Wang received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2019, and the Ph.D. degree from the School of Vehicle and Mobility, Tsinghua University, Beijing, China. He is currently with Rimian Kaiwu (Beijing) Technology Co., Ltd. His research interests include VLA/WAM post-training and real-world reinforcement learning for embodied intelligence.

Zhihuang Zhang received the B.E. degree in vehicle and mobility from Tsinghua University, Beijing, China, in 2018, and the Ph.D. degree in vehicle and mobility from Tsinghua University, Beijing, China, in 2023. His research interests include learning-based state estimation, time-series analysis, multi-sensor fusion, and high-precision localization for autonomous driving.

Mengmeng Yang received the Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University, Wuhan, China in 2018. She is currently an Associate Research Professor at Tsinghua University, Beijing, China. Her research interests include autonomous vehicles, high precision digital map, and sensor fusion.

Jin Huang received the B.E. and Ph.D. degrees from the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China, in 2006 and 2012, respectively. From 2009 to 2011, he was a joint Ph.D. with the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Since 2013 and 2016, he has been the Postdoc and an Assistant Research Professor with Tsinghua University, Beijing, China. He is currently an Associate Research Professor at Tsinghua University, Beijing, China. His research interests include artificial intelligence in intelligent transportation systems, dynamics control, and fuzzy engineering.

Zhihua Zhong received the Ph.D. degree in engineering from Linköping University, Linköping, Sweden, in 1988. He is currently a Professor with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. He was an Elected Member of the Chinese Academy of Engineering in 2005. His research interests include auto collision security technology, the punching and shaping technologies of the auto body, modularity and light-weighing auto technologies, and vehicle dynamics.

Diange Yang received the B.S. and Ph.D. degrees in automotive engineering from Tsinghua University, Beijing, China, in 1996 and 2001, respectively. He serves as the Director of automotive engineering at Tsinghua University. He is currently a Professor with the School of Vehicle and Mobility, Tsinghua University. His research interests include intelligent transport systems, vehicle electronics, and vehicle noise measurement. Dr. Yang attended in "Ten Thousand Talent Program" in 2016. He also received the Second Prize from the National Technology Invention Rewards of China in 2010 and the Award for Distinguished Young Science and Technology Talent of the China Automobile Industry in 2011.