2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

MapLocNet: Coarse-to-Fine Feature Registration for Visual Re-Localization in Navigation Maps

Hang Wu, Zhenghao Zhang, Siyuan Lin, Xiangru Mu, Qiang Zhao, Ming Yang, Tong Qin*

***Abstract*— Robust localization is the cornerstone of autonomous driving, especially in challenging urban environments where GPS signals suffer from multipath errors. Traditional localization approaches rely on high-definition (HD) maps, which consist of precisely annotated landmarks. However, building HD map is expensive and challenging to scale up. Given these limitations, leveraging navigation maps has emerged as a promising low-cost alternative for localization. Current approaches based on navigation maps can achieve highly accurate localization, but their complex matching strategies lead to unacceptable inference latency that fails to meet the real-time demands. To address these limitations, we introduce MapLocNet, a novel transformer-based neural re-localization method. Inspired by image registration, our approach performs a coarse-to-fine neural feature registration between navigation map features and visual bird's-eye view features. MapLocNet substantially outperforms the current state-of-the-art methods on both nuScenes and Argoverse datasets, demonstrating significant improvements in localization accuracy and inference speed across both single-view and surround-view input settings. We highlight that our research presents an HD-map-free localization method for autonomous driving, offering a cost-effective, reliable, and scalable solution for challenging urban environments.**

I. Introduction

With the recent advancements in autonomous driving technology, robust localization has emerged as a critical component. While both autonomous and human-driven vehicles rely heavily on Global Positioning System (GPS) for outdoor navigation, urban environments pose significant challenges to its accuracy. Multipath propagation errors and line-of-sight occlusions from urban structures often compromise GPS performance. In the absence of a reliable global localization source, positional accuracy can quickly deteriorate, leading to rapid location drift.

To survive in a GPS-denied situation, additional active localization methods are required. By leveraging prior built maps, such as 3D point clouds and distinctive visual features, LiDAR-based [1] and visual-based [2, 3] SLAM methods can be used for localization. However, this point-wise prior map is memory-consuming, and can not be used in the large environment for the autonomous driving task. Autonomous driving has relied heavily on high-definition (HD) maps containing precisely geo-referenced landmarks and geometries in the GPS-denied area. However, the exorbitant costs of producing and maintaining these maps have severely limited their scalability across diverse environments and geographies. As a result, the dependency on HD maps has been a major bottleneck preventing wider adoption of self-driving capabilities. With the advancement in perception algorithms, methods such as HDMapNet [4] and MapTR [5] have enabled online HD maps generation, allowing autonomous driving to be possible even with low-precision localization.

Tong Qin, Xiangru Mu, and Ming Yang are with Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China. Hang Wu, Zhenghao Zhang, Siyuan Lin, and Qiang Zhao are with IAS BU, Huawei Technologies, Shanghai, China. {qintong, muxiangru, mingyang}@sjtu.edu.cn {wuhang12, zhangzhenghao6, linsiyuan1, zhaoqiang20}@huawei.com. * is the corresponding author.

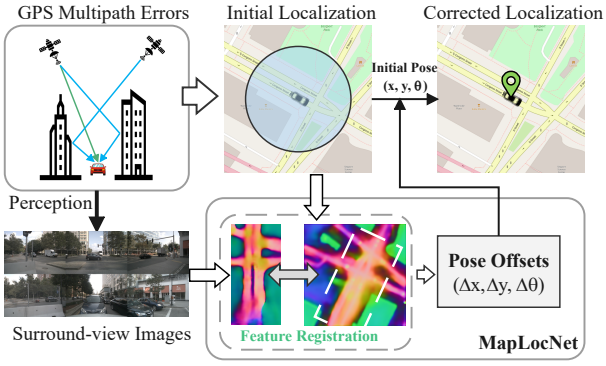


Fig. 1: Due to signal occlusion and multipath errors, the GPS-based positioning is unreliable in complex urban environments. To address this problem, we propose MapLocNet, which leverages surround-view images and navigation maps, utilizing consecutive neural localization modules based on coarse-to-fine feature registration principles to achieve superior localization accuracy in challenging scenarios. The video demonstration can be found at: https://youtu.be/3YG-Q5PBwOk.

On the other hand, we find some cues from biology, that human drivers can recognize location just with navigation maps. By correlating visual observations with map information, humans can roughly localize themselves in complex urban environments. Humans extract high-level semantics like road structures, building outlines, and landmarks from their surroundings, and leverage cognitive abilities to match these semantics with navigation maps. Currently, in the fields of robotics and augmented reality (AR), similar approaches [6, 7] have been proposed to emulate human-like localization methods. However, such methods often employ complex matching strategies for localization, making real-time inference unattainable. As a result, they cannot be applied in autonomous driving systems. To address the aforementioned challenges, we propose MapLocNet, a novel approach that achieves high localization accuracy while meeting real-time performance requirements. In our approach, we encode

surround-view images into the BEV space and process the navigation map using a U-Net [8]. The key innovation is the use of a transformer-based hierarchical feature registration method, which aligns visual BEV features with map features effectively, resulting in highly accurate localization. Our proposed approach surpasses current state-of-the-art (SOTA) methods in both localization accuracy and inference latency. Overall, the contribution of our paper is summarized as follows:

- We proposed MapLocNet, achieving highly accurate localization by fusing surround-view images and navigation maps, especially in GPS-denied areas suffered from significant positioning drift.
- We introduce a hierarchical coarse-to-fine feature registration strategy that aligns BEV and map features, attaining superior localization accuracy and inference speed compared to existing methods.
- We develop a novel training criterion that leverages perception tasks as auxiliary objectives for pose prediction. Our MapLocNet achieved the SOTA localization accuracy on the both nuScenes and Argoverse datasets.

We highlight that our research presents an HD-map-free, reliable and human-like localization approach, achieving superior localization accuracy compared to existing methods.

II. LITERATURE REVIEW

A. Localization Using Navigation Maps

Building the HD map was expensive, recent research focused on localization based on the lightweight navigation map. Panphattarasap *et al.* [9] proposed a novel approach to image-based localization in urban environments using semantic matching between images and a 2D map. Samano *et al.* [10] designed a novel method to geo-localize panoramic images on a 2D navigation map based on learning a low dimensional embedded space. Zhou *et al.* [11] presented a 2.5D map-based cross-view localization method that fuses the 2D image features and 2.5D maps to increase the distinctiveness of location embeddings. OrienterNet [6] proposed a deep neural network that estimates the pose of a query image by matching a neural BEV with available maps from OpenStreetMap (OSM) [12] and has achieved high-precision localization. Other methods [13]–[15] achieved cross-view geolocalization that matches the camera images from vehicles with an aerial image or a satellite image to determine the vehicle's pose. Drawing inspiration from the previous researches, we propose a localization approach that combines visual environmental perception with navigation maps.

B. BEV Representation for Visual Localization

There are many methods to transform the image features to the BEV grid, including geometry methods and learning-based methods. Cam2BEV [16] and VectorMapNet [17] used a geometry method that leveraged Inverse Perspective Mapping (IPM) to transform the image features to the BEV space through the plane assumption. HDMapNet [4] put forward a novel view transformer that consists of both neural feature extraction and geometric projection to get the BEV features. LSS [18], BEVDepth [19], BEVDet [20] learned a depth distribution of image features to lift each pixel to the 3D space. Then they used the camera extrinsic and intrinsic to splat all frustums into the BEV. GKT [21] proposed an efficient and robust 2D-to-BEV representation learning method that leveraged the geometric priors to guide the transformer to focus on discriminative regions, and unfolded kernel features to obtain BEV features. BEVFormer [22] leveraged predefined grid-shaped BEV queries to look up spatio-temporal space and aggregate spatio-temporal information from images, achieving SOTA performance on 3D object detection. To balance the precision and efficiency, we designed our BEV module based on the LSS architecture.

C. Image Registration

Image registration aims to find the spatial mapping between pixels in one image and another image, which is widely used in medical imaging and robotics research. Traditional feature-based methods [23, 24] leveraged the keypoints detected from images and its descriptors to match different images. Recently, CNN and transformer based image registration methods emerged to accelerate the registration time and accuracy. DIRNet [25] proposed a deep learning network for deformable image registration. The network included a ConvNet regressor, a spatial transformer, and a resampler. C2FViT [26] is a learning-based approach for 3D affine medical image registration that leverages the global connectivity of the self-attention mechanism and the locality of convolutional feed-forward layers to robustly encode global orientations and spatial relationships into a set of geometric transformation parameters. Following the principles of C2FViT, we construct a hierarchical feature registration module for visual localization.

D. End-to-end Localization Neural Networks

Recent end-to-end (E2E) approaches have introduced efficient architectures that estimate ego pose directly from sensor inputs and prior maps, circumventing the need for complex geometric calculations and hand-crafted rules. Using a differentiable optimization method, PixLoc [27] designed an E2E neural network to estimate the pose of an image by aligning deep features with a reference 3D model. I2D-Loc [28] proposed an effective network for camera localization based on the local image-LiDAR depth registration and used the BPnP module to calculate the gradients of the backend pose estimation for E2E training. BEV-Locator [29] designed a novel E2E architecture for visual semantic localization from multi-view images and a vectorized global map. Based on the cross-modal transformer structure, it addressed the key challenge of cross-modality matching between semantic map elements and camera images. EgoVM [30] built an E2E localization network that used light vectorized maps and has achieved centimeter-level localization accuracy. Inspired by the aforementioned works, our method builds an E2E localization network based on the transformer to achieve accurate localization.

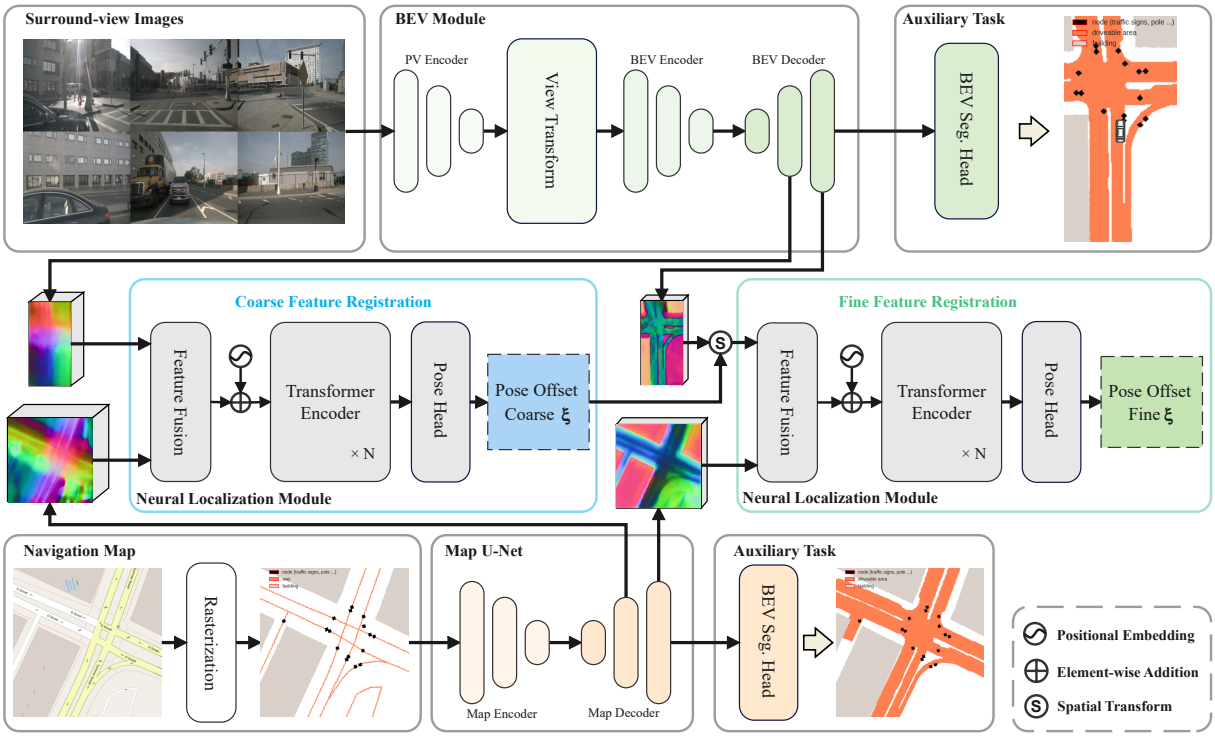


Fig. 2: The overall architecture of MapLocNet comprises three main modules: the BEV Module, Map U-Net, and Neural Localization Module. Our approach employs a coarse-to-fine feature registration strategy, extracting multi-scale features from both the BEV Decoder and Map Decoder to perform hierarchical feature alignment. Following the initial coarse registration stage, which yields a coarse estimate of the pose offset, we apply a spatial transformation to the high-resolution BEV features to facilitate the subsequent fine registration process. The predictions from both stages are combined to yield the final pose offset estimation result.

III. METHODOLOGY

A. Problem Formulation and System Overview

Given an initial vehicle localization $\check{\mathbf{p}} = (x, y, \theta)$ under noisy GPS, our goal is to estimate the transformation $\Delta\hat{\mathbf{T}} = \{\mathbf{R}, \mathbf{t}\}$ that transforms the initial noisy position to the ground-truth position $\mathbf{p}$. Since the re-localization is performed on a 2D navigation map, the pose transformation can be simplified to a 3-DoF transformation with a rotation $\mathbf{R} \in SO(2)$ and a 2D translation $\mathbf{t} \in \mathbb{R}^2$. The transformation can be solved by:

$$\min_{\Delta\hat{\mathbf{T}}} \left\| \Delta\hat{\mathbf{T}} \cdot \check{\mathbf{p}}, \mathbf{p} \right\|_2 . \tag{1}$$

Here, $\Delta\hat{\mathbf{T}}$ represents the transformation $\hat{\mathbf{T}}_{\mathbf{est}\leftarrow\mathbf{err}}$ from the erroneous localization to the estimated localization. We aim to minimize the difference between $\hat{\mathbf{T}}_{\mathbf{est}\leftarrow\mathbf{err}}$ and the ground-truth transformation $\mathbf{T}_{\mathbf{GT}\leftarrow\mathbf{err}}$. $\Delta\hat{\mathbf{T}}$ is denoted as $\hat{\boldsymbol{\xi}} = (\Delta x, \Delta y, \Delta\theta)$, where $(\Delta x, \Delta y) \in \mathbb{R}^2$ represent the longitudinal and lateral offsets, and $\Delta\theta \in (-\pi, \pi]$ represents the heading angle offset.

Our localization approach contains three key modules: the BEV Module, the Map U-Net and the Neural Localization Module. The overview of our network are shown in Fig. 2. The entire network is trained end-to-end using surround-view images, navigation maps, BEV segmentation, and ground-truth pose offsets. We leverage coarse and fine pose offsets regression and semantic segmentation losses during training to supervise consistent representations among visual BEV and map branchs. Details of the architecture and losses will be presented in the following sections.

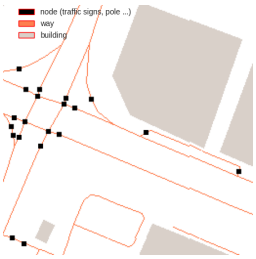
B. Map Processing

Since our method incorporates navigation map inputs and BEV semantic segmentation labels, it requires appropriate processing and fusion of different map data sources:

1) Map Rasterization: For accessibility and comprehensive coverage, we utilize OSM [12] as our navigation map data source, as depicted in Fig. 3a. OSM represents buildings using polygonal areas, roads using polylines, and traffic signals and other points of interest (PoI) using nodes. As shown in Fig. 3b, we retain only essential elements like buildings, roads, and PoIs (traffic signals, poles), whose spatial arrangement provides crucial geometric constraints for localization. For each query, we retrieve a patch of rasterized navigation map centered around the initial vehicle localization coordinates.

2) Segmentation Labels: The BEV semantic segmentation labels come from two sources. The drivable area labels are obtained from HD map data such as nuScenes [31]. As a complementary source, building and PoI labels are derived from the navigation map, such as OSM [32].

(a) Original Map



(b) Rasterized Map

(c) Surround-view Images

Fig. 3: Visualization of the original navigation map, its rasterized representation, and corresponding surround-view images. Rasterization enhances the expression of topological cues and spatial layout, emphasizing key elements like lane lines and building areas.

C. BEV Module

This module is designed to extract image features and project them into the BEV space to obtain BEV features. The visual input can be a monocular front-view image or multiple surround-view images. The more images used, the broader the perception range, resulting in improved localization accuracy. An example of surround-view images is illustrated as shown in Fig. 3c. We choose the simple yet effective LSS [18] architecture as the backbone.

We adopt the EfficientNet [33] as the Perspective-View (PV) Encoder to extract the image features $\mathbf{F_I}$. Following the LSS procedure, we combine the extrinsic and intrinsic parameters to project $\mathbf{F_I}$ to BEV space of size $H_{bev} \times W_{bev}$. We consider the longitudinal observation range to be wider than the lateral range, and thus set the spatial dimensions such that $H_{bev} = 2W_{bev}$. In different up-sampling stages of the BEV Decoder module, we extract the low-resolution, high-channel coarse features $\mathbf{F^c_{bev}} \in \mathbb{R}^{D_c \times \frac{1}{2}H_{bev} \times \frac{1}{2}W_{bev}}$ and the high-resolution, low-channel fine features $\mathbf{F^f_{bev}} \in \mathbb{R}^{D_f \times H_{bev} \times W_{bev}}$, which are used for subsequent two-stage coarse-to-fine feature registration. We supervise this module with a BEV semantic segmentation auxiliary task, which can better constrain the model's learning objective while also effectively improving localization accuracy.

D. Map U-Net

We adopt a U-Net architecture to extract features from the rasterized maps. To mitigate the modality gap between map features and visual BEV features, we innovatively introduce a BEV segmentation auxiliary task for this module. We employ a VGG-16 [34] backbone for encoding the map features. Similar to the BEV Module, at different stages of the Map Decoder, we also extract coarse and fine-level map features $\mathbf{F^c_{map}} \in \mathbb{R}^{D_c \times \frac{1}{2}H_{map} \times \frac{1}{2}W_{map}}$ and $\mathbf{F^f_{map}} \in \mathbb{R}^{D_f \times H_{map} \times W_{map}}$ for subsequent hierarchical feature registration. Here exists the relationship where $H_{map} = H_{bev}$ and $W_{map} = 2W_{bev}$. We utilize the same BEV segmentation labels as the BEV module to supervise this module, constraining the disparity between the two types of features.

E. Neural Localization Module

This module is responsible for the fusion of map and visual features as well as the decoding of pose offsets, making it the core module of MapLocNet. We designed various architectures for pose decoders, and through extensive experiments, determined coarse-to-fine feature registration as the final optimal solution. Other comparative approaches will be discussed in the section IV-B.

We formulate neural localization as a feature registration task. Inspired by C2FViT [26], we adopt a transformer encoder to perform self-attention computation on the fused visual BEV and map features in a coarse-to-fine manner. The modules for coarse and fine registration share the same architecture. As the widths of BEV features and map features are different, we zero-pad the BEV features in the width dimension to match the width of the map features. Considering the computing consuming, we downsampled the BEV and map features 4 times along the height and width dimension. Following the C2FViT, we also employ a 7×7 convolutional kernel to fuse them along channel dimension and flattened the fused features into sequential tokens for self-attention encoding of pose hidden features.

Since this is a pose-related task, positional encoding is crucial. We experimented with both learned and fixed positional encoding methods, eventually opting for sinusoidal positional encoding similar to that used in [35]. However, we shifted the origin of the positional encoding coordinates to the center of the feature map. The positional encoding is injected into the fused features through element-wise addition.

In each Neural Localization Module, we design N repeated transformer encoder layers, where in practice we set $N = 3$. Following this, there is a Pose Head consisting of 3 layers of MLP for pose decoding. The 3-DoF pose offset $\hat{\boldsymbol{\xi}}_{\boldsymbol{c}}$ estimated in the coarse feature registration stage is applied to the fine BEV features $\mathbf{F^f_{bev}}$. Subsequently, the spatially transformed BEV features, along with the fine map features $\mathbf{F^f_{map}}$, undergo fine feature registration to further narrow the gap from the ground truth pose and obtain $\hat{\boldsymbol{\xi}}_{\boldsymbol{f}}$. The cumulative outputs from both stages collectively serve as the final pose offset estimation.

F. Loss Function

During training, since we have the ground-truth localization $\mathbf{p}$, it is easier to simulate the noisy localization $\check{\mathbf{p}}$ by applying the transformation $\mathbf{T}$ to $\mathbf{p}$, where $\mathbf{T}$ is the inverse of the pose offset matrix from Eq. 1, denoted as $\mathbf{T_{err \leftarrow GT}}$, and simplified to the 3-variable form $\boldsymbol{\xi_1}$.

Approach	#Cams	Position Recall@Xm ↑				Orientation Recall@$X°$ ↑				#Param.(M)	GFLOPs ↓	FPS ↑
		$1m$	$2m$	$5m$	$10m$	$1°$	$2°$	$5°$	$10°$			
OrienterNet [6]	1	5.83	18.92	52.83	66.21	32.13	41.56	65.63	80.41	54.9	161.9	8.1
MapLocNet One-Stage (Ours)	1	7.14	22.39	60.48	83.90	35.68	60.40	87.49	95.31	20.7	**49.81**	**38.5**
MapLocNet (Ours)	1	**8.96**	**27.05**	**64.57**	**85.86**	**40.36**	**65.31**	**89.66**	**96.17**	22.9	52.15	23.8
U-BEV* [36]	6	16.89	41.60	71.33	83.46	–	–	–	–	–	–	–
MapLocNet DETR (Ours)	6	10.40	29.76	66.81	86.81	42.59	68.05	90.83	96.63	20.5	55.40	19.6
MapLocNet CA (Ours)	6	13.36	35.16	70.54	88.25	48.15	73.09	92.66	97.34	20.8	53.80	23.1
MapLocNet One-Stage (Ours)	6	16.32	40.56	74.27	89.69	53.71	78.13	94.49	98.05	20.7	**53.14**	**24.4**
MapLocNet (Ours)	6	**20.10**	**45.54**	**77.70**	**91.89**	**58.61**	**84.10**	**96.23**	**98.62**	22.9	56.15	18.2

TABLE I: Localization results on nuScenes dataset. The sign * denotes that the method only predicts 2-DoF pose offset (data w/o orientation noise). The inference speed in FPS for all models was measured on an NVIDIA V100 GPU.

As our method involves two stages, it can be considered as applying two successive transformations to the ground truth localization to obtain the noisy localization $\check{\mathbf{p}}$. Hence, the following relationship exists:

$$\hat{\mathbf{T}} = \hat{\mathbf{T}}_{\mathbf{2}} \cdot \hat{\mathbf{T}}_{\mathbf{1}}. \tag{2}$$

$\hat{\mathbf{T}}_{\mathbf{1}}$ is the prediction from the first stage, representing the transformation $\hat{\mathbf{T}}_{\mathbf{c}\leftarrow\mathbf{GT}}$ from the ground truth localization to the coarse localization, corresponding to $\hat{\boldsymbol{\xi}}_{\boldsymbol{c}}$. $\hat{\mathbf{T}}_{\mathbf{2}}$ represents the prediction from the second stage, representing the transformation $\hat{\mathbf{T}}_{\mathbf{f}\leftarrow\mathbf{c}}$ from the coarse estimated localization to the fine localization, corresponding to $\hat{\boldsymbol{\xi}}_{\boldsymbol{f}}$.

Based on the relationships above, the coarse pose offset $\hat{\boldsymbol{\xi}}_{\boldsymbol{c}}$ is supervised by ground truth $\boldsymbol{\xi}_{\mathbf{1}}$, with the Smooth L1 Loss:

$$\mathcal{L}_c = \left\| \hat{\boldsymbol{\xi}}_{\boldsymbol{c}}, \boldsymbol{\xi}_{\mathbf{1}} \right\|_{S1}, \tag{3}$$

where $\|\cdot\|_{S1}$ denotes the Smooth L1 Loss. The ground truth $\mathbf{T}_{\mathbf{2}}$ corresponding to $\hat{\mathbf{T}}_{\mathbf{2}}$ relies on the first stage prediction $\hat{\mathbf{T}}_{\mathbf{1}}$ and the ground truth $\mathbf{T}$, obtained from:

$$\mathbf{T}_{\mathbf{2}} = \mathbf{T} \cdot (\hat{\mathbf{T}}_{\mathbf{1}})^{-1}. \tag{4}$$

Denoting the simplified form of $\mathbf{T}_{\mathbf{2}}$ as $\boldsymbol{\xi}_{\mathbf{2}}$, it is used to supervise $\hat{\boldsymbol{\xi}}_{\boldsymbol{f}}$ also with a Smooth L1 Loss:

$$\mathcal{L}_f = \left\| \hat{\boldsymbol{\xi}}_{\boldsymbol{f}}, \boldsymbol{\xi}_{\mathbf{2}} \right\|_{S1}. \tag{5}$$

To better supervise the BEV module, we designed a BEV segmentation auxiliary task, following the approach outlined in LSS [18], to accelerate training convergence and improve model performance. We employ a Binary Cross-Entropy (BCE) loss to constrain the semantic segmentation output of the BEV module against the ground truth labels, resulting in $\mathcal{L}_{bev}$. On the other hand, to encourage the map features and visual BEV features to be as close as possible, thereby reducing the difficulty of feature registration, we also designed a BEV segmentation auxiliary task for the Map U-Net, with ground truth labels corresponding to map regions. Similarly, we adopted the BCE loss function to obtain $\mathcal{L}_{map}$.

The overall loss function is defined as:

$$\mathcal{L} = \lambda_c \cdot \mathcal{L}_c + \lambda_f \cdot \mathcal{L}_f + \lambda_{bev} \cdot \mathcal{L}_{bev} + \lambda_{map} \cdot \mathcal{L}_{map}, \tag{6}$$

where λ_c, λ_f, λ_{bev} and λ_{map} are weighting factors. During training, all weights are set to 1.

IV. Experiments

A. Experimental Settings

1) Dataset: Our proposed approach was trained and validated using two autonomous driving datasets, nuScenes [31] and Argoverse [37], to ensure a comprehensive evaluation. The nuScenes dataset comprises 1000 driving sequences captured in Boston and Singapore. We used the default train split of nuScenes, consisting of 850 sequences. The nuScenes validation set, comprising 150 sequences, served as our evaluation benchmark. The Argoverse dataset, featuring 113 scenes recorded in Miami and Pittsburgh, was utilized, with 65 scenes allocated to the training split and 24 scenes to the validation split. To address the absence of navigation map data in both nuScenes and Argoverse datasets, we enriched our dataset by acquiring navigation maps from OSM for the corresponding geographic regions. We aligned the navigation map with the HD map through localization coordinate transformation, following the methodology outlined in BLOS-BEV [38]. Fig. 3 visualizes the alignment of local section of the navigation map information and a frame from the nuScenes dataset at the same location.

2) Implementation Details:

a) Network Settings: The MapLocNet employs 6 surround-view images as the visual input unless otherwise specified. We utilize the EfficientNet-B0 [33] architecture as the image backbone, and all input images are resized to 128×352 resolution. During the training stage, we apply essential image data augmentations to enhance model robustness including random cropping, random flipping, and the one random drop of camera input. The ego vehicle's perceptual range in BEV space is defined as $[-64m, 64m]$ along the longitudinal axis and $[-32m, 32m]$ along the lateral axis, both with a resolution of 0.5 meters per pixel (mpp). The depth distribution is binned within the range $[4m, 60m]$ at a resolution of $1mpp$. For each frame, we take a $128m \times 128m$ patch from rasterized navigation map

Lanes	Buildings	Nodes	Longitudinal Recall@Xm ↑				Lateral Recall@Xm ↑				Orientation Recall@X° ↑			
			$1m$	$2m$	$5m$	$10m$	$1m$	$2m$	$5m$	$10m$	1°	2°	5°	10°
✓	✓	✓	32.08	54.20	81.97	93.34	48.71	71.91	90.66	97.11	58.61	84.10	96.23	98.62
✓	✓		28.07	48.75	77.64	90.79	41.48	64.73	88.00	96.54	51.44	75.74	93.61	97.67
✓			22.92	41.53	72.58	88.80	34.66	58.12	85.72	96.04	42.59	68.05	90.83	96.63

TABLE II: Localization results of various combination of input map elements on nuScenes dataset.

Pose Loss	BEV Loss	Map Loss	Longitudinal Recall@Xm ↑				Lateral Recall@Xm ↑				Orientation Recall@X° ↑			
			$1m$	$2m$	$5m$	$10m$	$1m$	$2m$	$5m$	$10m$	1°	2°	5°	10°
✓	✓	✓	32.08	54.20	81.97	93.34	48.71	71.91	90.66	97.11	58.61	84.10	96.23	98.62
✓	✓		27.21	47.62	76.89	90.49	40.48	63.73	87.69	96.52	50.01	74.54	93.20	97.53
✓			21.31	39.21	70.56	87.95	32.32	55.62	84.85	95.83	40.36	65.31	89.66	96.17

TABLE III: Localization results of various combination of loss functions on nuScenes dataset.

centered on the ego vehicle's position, with a resolution of $0.5mpp$.

b) Simulating 3-DoF GPS Error: We first align the pose and scale of rasterized navigation map with the HD map from nuScenes and Argoverse datasets. During training, we take a patch from the rasterized navigation map, centered on the ego vehicle's position. To simulate GPS errors, we apply random rotations $\theta \in [-30^\circ, 30^\circ]$ and translations $\mathbf{t} \in [-30m, 30m]$ to this patch. We then crop the central $128m \times 128m$ area as the biased map input of the MapLocNet.

c) Training Details: We train the model using 8 NVIDIA V100 GPUs for 200 epochs, which takes around 48 hours to converge. The model is optimized using an AdamW optimizer [39] with a weight decay of 1e-7, batch size of 8, and initial learning rate of 1e-4. We use a cosine annealing scheduler to adjust the learning rate during training.

B. Localization Results and Comparison

1) Comparison Methods:

a) OrienterNet: We employed the official implementation of OrienterNet [6] for training and evaluation on both nuScenes and Argoverse datasets. To ensure a fair comparison, given OrienterNet's limitation to monocular input, we conducted a parallel set of experiments using single-camera input for our method.

b) U-BEV: Since U-BEV's task is similar to ours, we directly referenced the data provided in its paper [36]. It's worth noting that its localization results do not include orientation predictions. Consequently, the initial localization may lack heading angle error perturbations, which simplifies the task to some extent. Given its utilization of 6 surround-view images, we categorize it within the reference group for 6-camera configurations.

c) MapLocNet DETR: We take inspiration from the decoder design in DETR [40], creatively treating the pose offset as the query $\mathbf{Q}$ to retrieve the fused features from the visual BEV features and map features. The feature fusion follows the approach described in section III-E. The features processed by the DETR decoder are then passed through the same 3-layer MLP Pose Head for pose decoding.

d) MapLocNet CA: Inspired by LoFTR [41] and GeoTransformer [42], we designed our neural localization module using a Cross-Attention (CA) module. We treat visual features as queries $\mathbf{Q}$ and map features as key $\mathbf{K}$ and value $\mathbf{V}$, enabling cross-domain attention computation. The resulting features are then decoded by the same Pose Head for pose estimation.

e) MapLocNet One-Stage: Our method is hierarchical, which to some extent affects inference speed. We want to investigate whether our method can meet usage requirements with only one stage, namely coarse feature registration, under constrained computational resources. Therefore, we tested a one-stage version of MapLocNet here. To minimize computational complexity, we employed coarse features instead of fine features for our one-stage experiments. The sole difference from the hierarchical version is that we omitted fine feature registration and directly used the output of the initial coarse stage as the final result. We expect the one-stage version to balance localization accuracy and inference speed, while the coarse-to-fine version can push the upper bound of localization accuracy.

2) Localization Results:

a) nuScenes: For simplicity, we collectively term both one-stage and coarse-to-fine approaches as feature registration (FR) architecture. As shown in Tab. I, among the experimental groups with 6 cameras, the coarse-to-fine FR architecture achieved the best localization performance. Our one-stage FR architecture demonstrated the highest efficiency, achieving 24.4 frames per second (FPS). In the monocular experimental group, we compared the localization performance of the FR architecture with OrienterNet. Our approach surpassed OrienterNet in both computational efficiency and accuracy, notably exceeding OrienterNet's speed by approximately 30 FPS.

b) Argoverse: To further demonstrate the capabilities of our model, we conducted experiments on the Argoverse [37]

Approach	Position Recall@Xm ↑				Orientation Recall@$X°$ ↑			
	$1m$	$2m$	$5m$	$10m$	$1°$	$2°$	$5°$	$10°$
O-1	8.56	21.20	54.90	72.16	18.72	31.07	63.05	81.97
M-1	9.12	27.61	66.31	88.71	41.22	66.54	90.32	96.53
M-6	**23.26**	**47.24**	**79.13**	**94.33**	**62.35**	**86.28**	**96.24**	**98.61**

TABLE IV: Localization results on Argoverse dataset. Specifically, approach *O-1* represents OrienterNet [6] with single-view input, while *M-1* and *M-6* denote MapLocNet with single-view and six-view inputs, respectively.

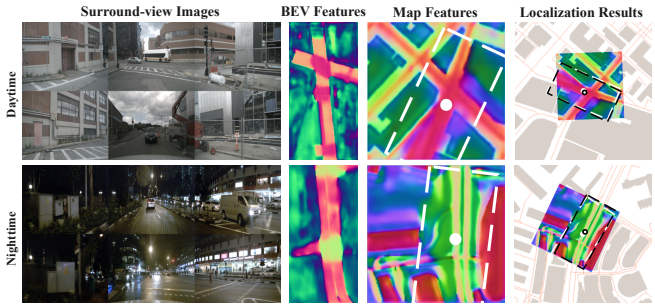


Fig. 4: Visualization of localization results in the nuScenes dataset during day and night. The middle two columns depict high-resolution, low-channel BEV features and map features, respectively. The white dots and bounding boxes in the map features represent the GT locations and orientations of the BEV features. In the last column, black dots and bounding boxes represent the corrected locations and orientations on the map after applying the offsets predicted by the model.

dataset. Leveraging pre-trained weights obtained from the nuScenes dataset, we fine-tuned our model on the Argoverse dataset. We also compared our approach with OrienterNet, employing identical training strategies. Remarkably, as depicted in Tab. IV, our model exhibits superior localization performance across both single-view camera and surround-view camera configurations. It significantly outperforms OrienterNet in accuracy across all input settings, highlighting the robustness and versatility of our approach.

3) Results Visualization: To illustrate the model's performance intuitively, Fig. 4 displays only the high-resolution, low-channel BEV features and map features utilized in the second-stage registration. The high-dimensional features employed in the initial coarse registration are omitted due to their visual complexity. Our experiments reveal a minor performance degradation in night scenes, attributed to reduced building visibility. Nevertheless, the model exhibits robust localization capabilities across both daytime and nighttime conditions.

C. Ablation Study

We performed comprehensive ablation studies to evaluate the influence of various OSM element combinations and loss function configurations on model performance.

1) Input OSM Elements: We conducted an ablation study on three key map elements: lanes, buildings, and nodes (including traffic lights and signs). Given their environmental prevalence, we sequentially removed nodes and buildings from the input. Tab. II demonstrates that all three elements contribute positively to localization performance. The removal of buildings caused a more significant performance drop compared to nodes, indicating their greater impact on localization. Notably, the model maintained substantial performance with lanes alone, suggesting their crucial role in localization. We posit that from a BEV perspective, the learning complexity decreases from nodes to buildings to lanes, while their environmental prevalence increases. This correlation aligns with their increasing importance in localization performance.

2) Loss Functions: This experiment investigates the impact of auxiliary segmentation tasks on localization performance. We introduce the BEV Loss, which guides feature learning in the visual branch, and the Map Loss, which uses the same semantic labels to reduce feature modality differences between visual and map branches. As shown in Tab. III, incorporating visual BEV segmentation loss supervision substantially enhances the model's localization performance. We posit that this loss improves the model's comprehension of environmental structures, providing clearer localization cues. After adding map segmentation loss supervision, the model's localization performance is further improved. We hypothesize that a modality gap exists between the rasterized map and visual BEV representations. By unifying semantic supervision across both branches, we mitigate this modality discrepancy, thereby improving the model's localization capabilities.

V. Conclusion

In this work, we propose MapLocNet, a novel approach that achieves highly accurate and reliable localization by integrating surround-view images and navigation maps. Our method effectively addresses the challenge of re-localization under significant positioning drift, particularly in complex urban environments. Through hierarchical feature registration, our approach surpasses existing methods in both localization accuracy and inference speed on both the nuScenes and Argoverse datasets. The two-stage feature registration ensures high precision, while the one-stage approach balances accuracy and speed. Supervised BEV segmentation in both BEV and map modules mitigates modality gap, enhancing localization capability. Our method represents a notable advancement towards robust and precise vehicle localization, facilitating safe and reliable autonomous driving.

Limitations and Future Work: Our method currently uses single-frame inference and relies on BEV semantic supervision. Future work will focus on incorporating multi-frame information, exploring vectorized navigation map representations, and developing methods to mitigate multi-modal discrepancies without semantic supervision. These advancements aim to enhance localization accuracy, reduce computational overhead, and improve overall performance.

REFERENCES

[1] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9, 2014, pp. 1–9.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

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

[5] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," in *International Conference on Learning Representations*, 2023.

[6] P.-E. Sarlin, D. DeTone, T.-Y. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulò, R. Newcombe, P. Kontschieder, and V. Balntas, "Orienternet: Visual localization in 2d public maps with neural matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 632–21 642.

[7] P.-E. Sarlin, E. Trulls, M. Pollefeys, J. Hosang, and S. Lynen, "SNAP: Self-Supervised Neural Maps for Visual Positioning and Semantic Understanding," in *NeurIPS*, 2023.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer*, 2015, pp. 234–241.

[9] P. Panphattarasap and A. Calway, "Automated map reading: image based localisation in 2-d maps using binary semantic descriptors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6341–6348.

[10] N. Samano, M. Zhou, and A. Calway, "You are here: Geolocation by embedding maps and images," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 502–518.

[11] M. Zhou, L. Liu, and Y. Zhong, "Image-based geolocalization by ground-to-2.5 d map matching," *arXiv preprint arXiv:2308.05993*, 2023.

[12] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.

[13] T. Lentsch, Z. Xia, H. Caesar, and J. F. Kooij, "Slicematch: Geometry-guided aggregation for cross-view pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 225–17 234.

[14] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, and R. Stiefelhagen, "Uncertainty-aware vision-based metric cross-view geolocalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 621–21 631.

[15] Z. Xia, O. Booij, M. Manfredi, and J. F. Kooij, "Visual cross-view metric localization with dense uncertainty estimates," in *European Conference on Computer Vision*. Springer, 2022, pp. 90–106.

[16] L. Reiher, B. Lampe, and L. Eckstein, "A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[17] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.

[18] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[19] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[20] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[21] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, and W. Liu, "Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer," *arXiv preprint arXiv:2206.04584*, 2022.

[22] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[23] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[25] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Isgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," *CoRR*, vol. abs/1704.06065, 2017. [Online]. Available: http://arxiv.org/abs/1704.06065

[26] T. C. Mok and A. Chung, "Affine medical image registration with coarse-to-fine vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 835–20 844.

[27] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-perfect structure-from-motion with featuremetric refinement," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5987–5997.

[28] K. Chen, H. Yu, W. Yang, L. Yu, S. Scherer, and G.-S. Xia, "I2d-loc: Camera localization via image to lidar depth flow," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 194, pp. 209–221, 2022.

[29] Z. Zhang, M. Xu, W. Zhou, T. Peng, L. Li, and S. Poslad, "Bev-locator: An end-to-end visual semantic localization network using multi-view images," *arXiv preprint arXiv:2211.14927*, 2022.

[30] Y. He, S. Liang, X. Rui, C. Cai, and G. Wan, "Egovm: Achieving precise ego-localization using lightweight vectorized maps," *arXiv preprint arXiv:2307.08991*, 2023.

[31] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[32] P. Mooney, M. Minghini *et al.*, "A review of openstreetmap data," *Mapping and the citizen sensor*, pp. 37–59, 2017.

[33] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[36] A. B. Camiletto, A. Bochicchio, A. Liniger, D. Dai, and A. Gawel, "U-bev: Height-aware bird's-eye-view segmentation and neural map-based relocalization," *arXiv preprint arXiv:2310.13766*, 2023.

[37] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[38] H. Wu, Z. Zhang, S. Lin, T. Qin, J. Pan, Q. Zhao, C. Xu, and M. Yang, "Blos-bev: Navigation map enhanced lane segmentation network, beyond line of sight," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 3212–3219.

[39] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv*, vol. abs/1711.05101, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3312944

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *ArXiv*, vol. abs/2005.12872, 2020.

[41] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," *CVPR*, 2021.

[42] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 143–11 152.