

Article

NMLoNet: An End-to-End Intelligent Vehicle Localization Network Using Navigation Maps

Qingtong Yuan * and Yicheng Li

Automotive Engineering Research Institute (AERI), Jiangsu University, Zhenjiang 212013, China;
liyicheng070@ujs.edu.cn

* Correspondence: 2212338016@stmail.ujs.edu.cn

Abstract

Accurate and reliable localization is crucial for advanced autonomous driving systems. Traditional high-precision localization approaches rely on meticulously annotated high-definition (HD) maps and employ visual-geometric methods to derive accurate pose information. However, the construction, maintenance, and updating of HD maps are costly and time-consuming. In contrast, localization using publicly available navigation maps provides a low-cost and scalable alternative. Existing methods typically align BEV (Bird's-Eye-View) features extracted from surround-view images with navigation maps to obtain localization results. Although such approaches can achieve high accuracy, they often neglect the inherent modality gap between BEV features and navigation maps, leading to localization errors. To address this issue, we propose NMLoNet: An End-to-End Intelligent Vehicle Localization Network Using Navigation Maps. The proposed method exploits road semantic elements to effectively bridge the modality gap between BEV representations and navigation maps. Specifically, a Deformable Attention Module is introduced after BEV feature extraction to capture long-range dependencies among BEV features. Furthermore, we innovatively incorporate vector map constraints to minimize the discrepancy between BEV and navigation map features. In addition, a multi-level cross-modal feature registration mechanism is designed to achieve more precise alignment between BEV and map representations. Extensive experiments on the nuScenes and Argoverse datasets demonstrate that NMLoNet achieves state-of-the-art performance, improving localization accuracy by approximately 11% under monocular settings and 24% under surround-view configurations. Moreover, the proposed network maintains robust localization performance in complex and highly dynamic driving environments.

Keywords: visual localization; navigation map; bird's-eye-view; intelligent vehicles; high-precision localization



Academic Editor: Kan Akatsu

Received: 10 January 2026

Revised: 25 February 2026

Accepted: 27 February 2026

Published: 17 March 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the World Electric Vehicle Association.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

With the rapid development of modern technology, automobiles have become an indispensable means of transportation in daily life. However, this growing reliance on vehicles has also led to severe challenges such as traffic congestion and frequent accidents. The intelligentization of vehicles has been recognized as an effective approach to addressing these issues [1]. Moreover, as people's living standards improve, the demand for intelligent vehicles continues to increase, making vehicle intelligence a major focus of current research and industrial development.

Accurate and robust localization is fundamental to intelligent driving systems, as downstream tasks—including motion planning, decision-making, and vehicle control—depend critically on precise vehicle pose estimation [2]. In outdoor environments, autonomous vehicles typically rely on the Global Navigation Satellite System (GNSS) for localization. However, GNSS suffers from low update rates and limited positioning accuracy, which are insufficient for high-precision autonomous driving. Moreover, GNSS signals are vulnerable to occlusions caused by urban structures such as trees, high-rise buildings, and overpasses. This leads to the well-known urban canyon effect, resulting in severe positioning degradation or even complete signal loss [3]. To alleviate these limitations, GNSS is often integrated with an Inertial Navigation System (INS). In such systems, GNSS compensates for long-term INS drift, while INS mitigates the low update frequency of GNSS. Nevertheless, when GNSS signals become unavailable, the integrated system loses its global reference, causing rapid drift accumulation and significant localization errors. Due to the inherent uncertainty of GNSS- and INS-based solutions, additional reliable localization methods are required. Simultaneous Localization and Mapping (SLAM) is one of the most widely adopted active localization approaches. Camera-based SLAM methods [4,5] are sensitive to illumination changes and environmental variations, which limits their robustness. In contrast, LiDAR-based SLAM methods [6,7] provide higher accuracy and stronger environmental resilience. However, their large-scale deployment remains challenging due to the high cost of LiDAR (Light Detection and Ranging) sensors. Currently, high-definition (HD) map-based localization is the dominant solution in autonomous driving [8–10]. HD maps are typically constructed using high-channel LiDAR systems or complex multi-camera reconstruction pipelines, encoding dense static landmarks with centimeter-level accuracy. Vehicles achieve precise localization by aligning real-time perception data with HD map landmarks, usually through geometric techniques such as Perspective-n-Point (PnP) and RANSAC. Despite their effectiveness, HD maps are expensive to construct, maintain, and update, which significantly limits scalability and large-scale deployment. Furthermore, traditional geometric alignment methods remain sensitive to dynamic environments and visual degradation. These limitations have motivated the search for low-cost, scalable, and robust alternatives to HD map-based localization.

For humans, determining one's position using publicly available navigation maps is intuitive. For autonomous systems, however, this task is considerably more challenging. Compared with HD maps, navigation maps lack dense point cloud information and detailed geometric descriptors. Instead, they primarily contain simplified structural elements such as road boundaries, building outlines, and lane-level topology. As a result, direct geometric alignment between camera observations and navigation maps is infeasible. Recent advances in computer vision have opened new possibilities. Online HD map generation frameworks such as HDMapNet [11] and MapTR [12] demonstrate that high-level semantic vector representations can be extracted directly from visual inputs. These semantic vector elements exhibit strong structural consistency with navigation maps, suggesting that localization can be achieved by aligning online-generated semantic vectors with navigation map features. Prior studies [13,14] explored related ideas in augmented reality (AR) for indoor environments. However, these approaches are not designed for the complexity of autonomous driving scenarios. MapLocNet [15] was the first framework to leverage navigation maps for vehicle localization. Nevertheless, it overlooked the modality discrepancy between Bird's-Eye-View (BEV) features derived from surround-view images and 2D navigation map representations. This cross-modal domain gap significantly degrades alignment performance, particularly in dynamic or repetitive environments, leading to substantial localization errors. In this work, we explicitly address this modality gap. We introduce an online vector map constraint after BEV feature extraction to guide the BEV

representation toward the semantic distribution of navigation map features. In addition, we design a BEV Feature Enhancement Module that strengthens both semantic and geometric representations through high-order feature modeling. This module enables the encoder to adaptively emphasize localization-relevant road elements. Furthermore, we develop a multi-level cross-modal fusion mechanism to achieve hierarchical and comprehensive alignment between BEV and navigation map features. In summary, the main contributions of this paper are as follows:

- (1) To address the problem that autonomous driving localization heavily relies on expensive high-definition (HD) maps, we propose NMLoNet, an end-to-end localization network based on surround-view images and navigation maps. Without using costly HD maps, NMLoNet achieves autonomous and reliable localization by leveraging publicly available navigation maps.
- (2) To mitigate the modality gap between BEV features and 2D navigation map features, we innovatively introduce an online vector map constraint, which effectively reduces the discrepancy between the two modalities. In addition, to address the discreteness of BEV features and their limited ability to continuously and efficiently focus on road-related localization elements, we design a BEV feature enhancement module. This module captures the correlations among BEV features through adaptive weighting, enabling the network to continuously and efficiently focus on road localization elements.
- (3) To further tackle the difficulty of aligning BEV features with navigation map features, we propose BMF, a feature alignment module proposed in this paper, which adaptively learns both feature positions and feature weights. Experimental results demonstrate that our NMLoNet achieves state-of-the-art performance on the nuScenes and Argoverse datasets.

2. Related Work

2.1. Vision-Based Geometric Localization

In recent years, visual geometry-based localization methods have dominated intelligent vehicle positioning. These approaches estimate vehicle poses by matching environment descriptors or LiDAR point clouds from the current scene with pre-built 3D maps. Representative paradigms include Simultaneous Localization and Mapping (SLAM) and localization based on pre-constructed high-definition (HD) maps. SLAM techniques have been extensively studied over the past decades, and numerous mature frameworks [4–7,16,17] have been widely adopted in both academia and industry. Representative visual SLAM systems such as ORB-SLAM [5,16] and VINS [4] are broadly used in augmented reality (AR) and robotics due to their low hardware requirements and lightweight memory footprint. However, their limited precision and sensitivity to illumination and environmental variations restrict their applicability in autonomous driving scenarios. In contrast, LiDAR-based SLAM approaches [7,17,18] provide higher accuracy and stronger robustness, which makes them more suitable for autonomous driving. Nevertheless, their large-scale deployment remains constrained by the high cost of LiDAR sensors and substantial computational demands. More importantly, both vision-based and LiDAR-based SLAM methods suffer from accumulated drift errors. In large-scale driving environments, such drift can significantly degrade localization accuracy, posing serious risks for safety-critical applications.

Localization approaches based on pre-built HD maps [8,19–23] effectively mitigate drift accumulation and are therefore more appropriate for large-scale autonomous driving. In these systems, real-time sensor data—such as point clouds or visual descriptors—are aligned with HD maps using geometric algorithms, including Perspective-n-Point (PnP) and Efficient PnP (EPnP), to estimate vehicle poses. However, HD map-based localization

also faces practical challenges. In open outdoor environments, the large scale of HD maps makes efficient descriptor matching computationally demanding, often compromising real-time performance and even leading to localization failure. To improve efficiency, refs. [9,19] exploited structural landmarks such as poles and road curbs to simplify registration. In [8,20], the environment was partitioned into node maps associated with GNSS signals, enabling a coarse-to-fine localization strategy that improved matching success rates. A crowdsourced lightweight semantic mapping method was proposed in [21] to reduce mapping costs, while [24] introduced semantic keyframes into node maps to support efficient monocular localization. Additionally, ref. [25] extracted pole features via instance segmentation and employed Gaussian regression to estimate their relative pose with respect to the map. Despite their high accuracy, HD map-based localization methods incur substantial construction, maintenance, and storage costs, as maps must be continuously updated to reflect environmental changes. Furthermore, both SLAM and HD map-based approaches rely heavily on visual geometric optimization, which inherently introduces uncertainty. When deployed at a large scale, such uncertainties may result in non-negligible random localization errors, ultimately undermining system reliability.

2.2. End-to-End Localization Neural Network

With the rapid advancement of computer vision, end-to-end (E2E) localization frameworks have emerged as a prominent research direction. These approaches directly regress vehicle poses from raw sensor inputs, thereby avoiding the complexity and potential instability of traditional visual-geometric optimization pipelines. PoseNet [23] was the first to introduce a framework that directly regresses camera poses from a single RGB image by appending a multilayer perceptron (MLP) to an image feature extractor backbone (GoogLeNet). However, due to the absence of map-based constraints, PoseNet typically produces meter-level localization errors, which are insufficient for intelligent driving applications. Subsequent works [10,25–27] improved localization accuracy to some extent, yet still failed to achieve the centimeter-level precision required for autonomous vehicles. To incorporate stronger structural priors, L3-Net [28] aligns real-time LiDAR keypoints with pre-built HD maps through a 3D estimation volume, enabling end-to-end training while achieving centimeter-level accuracy. Similarly, ref. [29] leverages stable features extracted from road-invariant structures to improve localization robustness. In [30], lane markings, traffic signs, and sparse semantic maps are utilized for pose estimation, effectively reducing map storage requirements. PixLoc [31] aligns deep image features with corresponding 3D models in a fully differentiable framework, allowing joint optimization of feature extraction and pose regression. In the aforementioned studies, inputs are typically real-time point clouds or image-derived depth representations. With the rise of multimodal perception, camera-based localization frameworks have been further developed to exploit visual information more effectively.

CMRNet [32] constructs a cost volume between image features and LiDAR map features using optical flow estimation, followed by a regression module to predict the vehicle's relative pose. Building upon this idea, ref. [33] introduces a Transformer-based architecture to model long-range correspondences between 2D images and 3D LiDAR maps, achieving improved localization accuracy. In [34], point cloud maps are compressed and decoded into virtual feature images, which are then matched with camera images to estimate relative poses. Despite these advancements, the substantial modality gap between 2D monocular images and 3D LiDAR maps remains a major challenge, making accurate pose regression from single images difficult. To address this issue, BEV-Locator [35] was the first to leverage multi-view images and global vector maps for vision-based semantic localization. By employing a cross-modal Transformer architecture, it effectively bridges the

gap between navigation maps and camera-derived features, achieving centimeter-level accuracy. Similarly, EgoVM [36] integrates learnable semantic embeddings and maps element coordinates into BEV representations to enhance cross-modal alignment and further adopts a histogram-based optimization strategy for robust pose estimation. Inspired by these end-to-end localization paradigms, our method follows a fully differentiable framework while introducing novel architectural components to further improve localization accuracy and robustness.

2.3. Localization with Navigation Maps

Due to the high cost and logistical complexity of constructing and maintaining high-definition (HD) maps, recent research has shifted toward HD-map-free localization for intelligent vehicles. One line of work leverages aerial imagery for cross-view localization by matching ground-level camera images with overhead views [37–40]. However, aerial imagery is often unavailable or restricted due to environmental conditions (e.g., cloud cover), privacy regulations, and limited coverage—severely limiting the practicality and scalability of such approaches. To address this, an increasing number of studies have explored localization using lightweight, publicly available navigation maps (e.g., OpenStreetMap). For example, ref. [41] performs semantic matching between monocular images and 2D navigation maps to estimate vehicle pose. In [42], a geolocalization method is proposed that embeds panoramic images and navigation maps into a shared low-dimensional space for efficient retrieval. Similarly, ref. [43] achieves high-precision localization by matching LiDAR-derived descriptors with those extracted from 2D maps, attaining accuracy comparable to HD-map-based systems. More recently, ref. [44] combines human-perspective and BEV (Bird’s-Eye View) images for map matching against navigation maps, demonstrating promising results in indoor settings. A major breakthrough was introduced by OrienterNet [14], which first established end-to-end localization by aligning BEV features—extracted from surround-view cameras—with vectorized OpenStreetMap (OSM) data [45]. This idea was extended by MapLocNet [15], which formulates localization as an image registration problem between BEV feature maps and OSM raster maps. In contrast to these methods, which treat BEV-to-map alignment as a pixel-level image registration task, we argue that this formulation overlooks the fundamental modality gap: BEV features are not raw images but high-level, semantically rich representations encoding road topology, drivable areas, and spatial context. Directly registering them with geometric navigation maps—as if they were conventional images—leads to suboptimal alignment and information loss. Therefore, in this paper, we reformulate the problem as a cross-modal feature registration task, where the goal is to semantically align and adaptively fuse heterogeneous representations—BEV visual semantics and map geometry—through a dedicated neural architecture. This perspective enables more precise, robust, and interpretable localization without reliance on HD maps.

3. Materials and Methods

This section presents the implementation details of NMLoNet. NMLoNet is shown in Figure 1. NMLoNet is an end-to-end visual localization neural network that estimates the vehicle pose using surround-view images and navigation maps. The localization task can be formulated as follows: We consider a single frame of surround-view images with a noisy prior pose $\hat{p} = (x, y, \theta)$, which can be obtained from GPS or other consumer-grade sensors. Our goal is to estimate the optimal pose transformation $\hat{\zeta} = (\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{\theta})$ on a 2D navigation map, thereby transforming the initial noisy pose $\hat{p}$ into the true vehicle pose p . Here, where $(\Delta x, \Delta y) \in \mathbb{R}^2$ denote the longitudinal and lateral offsets, and $\Delta\theta \in (-\pi, \pi]$

represents the heading angle offset. The objective of our localization problem can thus be formulated as:

$$\min_{(\Delta x, \Delta y, \Delta \theta)} \|\hat{p} + \hat{\zeta}, p\|_2 \quad (1)$$

where $\|\cdot\|_2$ denotes the L2 Euclidean norm.

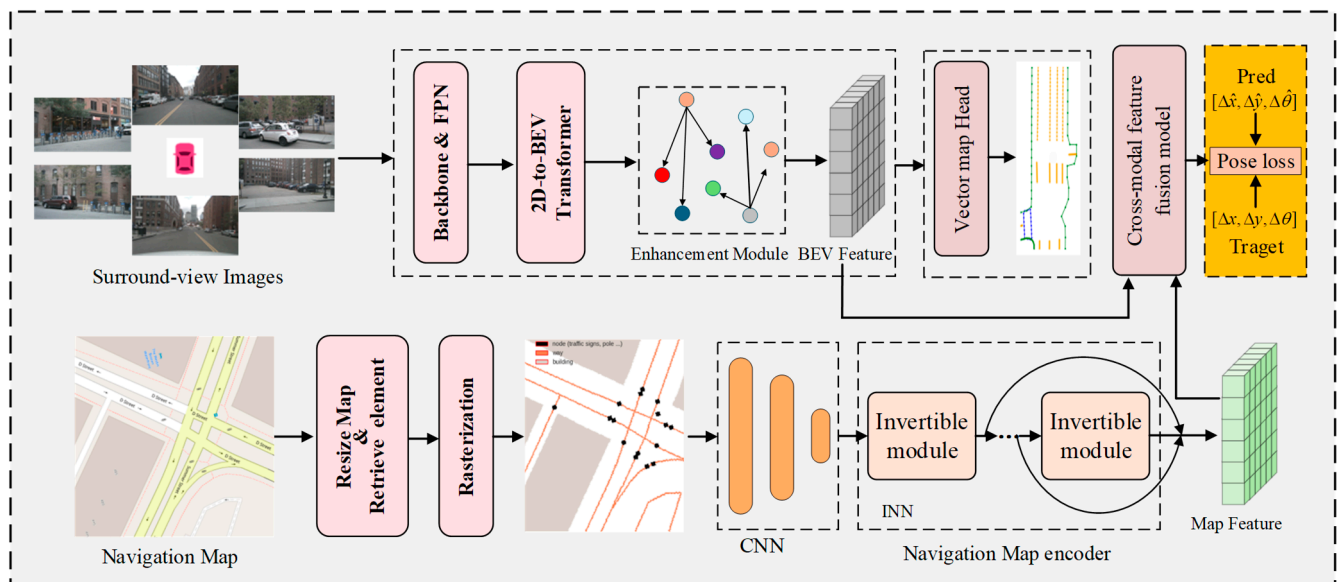


Figure 1. The overall framework of our proposed method is illustrated above. The inputs include six surround-view images centered on the ego vehicle and an SD navigation map containing road elements. The six images are transformed into BEV features through a view transformation module, while the 2D navigation map is encoded into SD features via a CNN-based encoder. In addition, the BEV features are supervised by an online map constraint module. Finally, the BEV and SD features are fused to output the vehicle pose.

3.1. BEV Feature Encoder

The purpose of this module is to project the surround-view image features from multiple cameras into the BEV (Bird's Eye View) feature space to obtain BEV representations. In this paper, we adopt the simple and efficient LSS framework [46] as the BEV backbone to enhance localization real-time performance. Meanwhile, NMLoNet is also compatible with other BEV feature extraction algorithms, such as GKT and BEVFormer [47]. We employ a pre-trained EfficientNet [48] as the backbone for surround-view visual feature extraction to obtain image features $F_{I(I=1,2,3 \dots 6)}$. Subsequently, the image features F_I are projected into BEV features $F_B \in R^{H \times W \times C}$, where C denotes the number of channels, and H and W represent the height and width of the visual BEV.

3.2. BEV Enhancement Module

This paper aims to reduce the modality gap between visual BEV features and 2D navigation map representations to improve localization accuracy. Compared with conventional street-view images, navigation map images exhibit significantly lower visual complexity. Their limited appearance information makes cross-modal registration particularly challenging. However, navigation maps are primarily composed of structured road vector elements. This observation motivates us to leverage real-time road perception and extract corresponding road vector features from BEV representations, thereby narrowing the modality gap between BEV features and navigation map images. In end-to-end vectorized HD map construction tasks [11,12], the road vector elements encoded in BEV features demonstrate

strong structural consistency with those in navigation maps. Inspired by this property, we introduce an auxiliary end-to-end vector map construction task to supervise BEV feature learning. This auxiliary supervision encourages BEV representations to emphasize road vector elements, effectively reducing the domain discrepancy between visual BEV features and navigation map representations. To further refine BEV features, we design a dedicated BEV Feature Enhancement Module. Existing studies have primarily focused on accurately transforming perspective-view images into bird's-eye-view representations. However, BEV features themselves contain rich structural and semantic information, and their internal dependencies deserve more careful modeling. In vectorized mapping tasks, most prior approaches emphasize modeling relationships among points within the same instance while overlooking strong inter-instance correlations and complex geometric relationships across different map elements. As a result, subtle structural variations and intricate spatial dependencies are often insufficiently captured. In our setting, the goal is not to generate precise vector maps but to extract robust BEV road semantic features for cross-modal alignment. Ignoring inter-element correlations and geometric dependencies would lead to fragmented feature representations, ultimately degrading alignment performance with 2D navigation map features. To address this issue, we introduce a Transformer-based self-attention mechanism to enhance BEV feature representation. Self-attention enables the modeling of complex geometric relationships across instances and captures fine-grained semantic variations among road elements. It also helps emphasize localization-relevant cues while suppressing irrelevant information. Nevertheless, conventional Transformer backbones [49–52] are computationally expensive and often exhibit slow convergence, making them unsuitable for real-time localization systems. Moreover, in vectorized mapping scenarios, dependencies among map elements are typically localized rather than globally distributed. Applying full global attention may introduce unnecessary noise and negatively affect BEV feature quality. Therefore, we adopt the Deformable Attention Module from Deformable DETR [53] as our BEV Feature Enhancement Module. This mechanism enables efficient and spatially adaptive attention computation, capturing local structural dependencies while maintaining computational efficiency. As a result, BEV feature representation is significantly enhanced without compromising real-time performance.

Specifically, as illustrated in Figure 2, the input BEV feature is denoted as $F_B \in R^{H \times W \times C}$, where C represents the number of channels, and H and W denote the height and width of the BEV feature, respectively.

First, the BEV features are linearly projected to obtain query tokens $q = F_B W_q$ for each pixel. Then, a set of reference points $p_q \in R^{H \times W \times 2}$ forming a uniform grid is generated, where each reference point $\{(0, 0), \dots, (H, W)\}$ is a 2D coordinate with linearly spaced values. The deformable attention feature is computed as follows:

$$\text{DeformAtten}(q, p_q, F_B) = \sum_{m=1}^M M_m \sum_{k=1}^K (W_{mqk} \cdot V_m F_B(p_q + \Delta p_{mqk})) \quad (2)$$

Here, m denotes the number of attention heads, k represents the sampled keys, and K represents the total number of sampled keys. M_m is the weight of each attention head. M_m is generated by a linear layer. $V_m F_B$ corresponds to the v in the traditional Transformer, $\Delta p_{mqk} \in R^2$ and $W_{mqk} \in [0, 1]$ ($\sum_{k=1}^K A_{mqk} = 1$) represent the offset and attention weight of the k -th sampling point in the m -th attention head, respectively. Both are obtained through fully connected layers.

$$V_m F_B = Fc1(q) \quad \Delta p_{mqk} = Fc2(q) \quad W_{mqk} = Fc3(q) \quad (3)$$

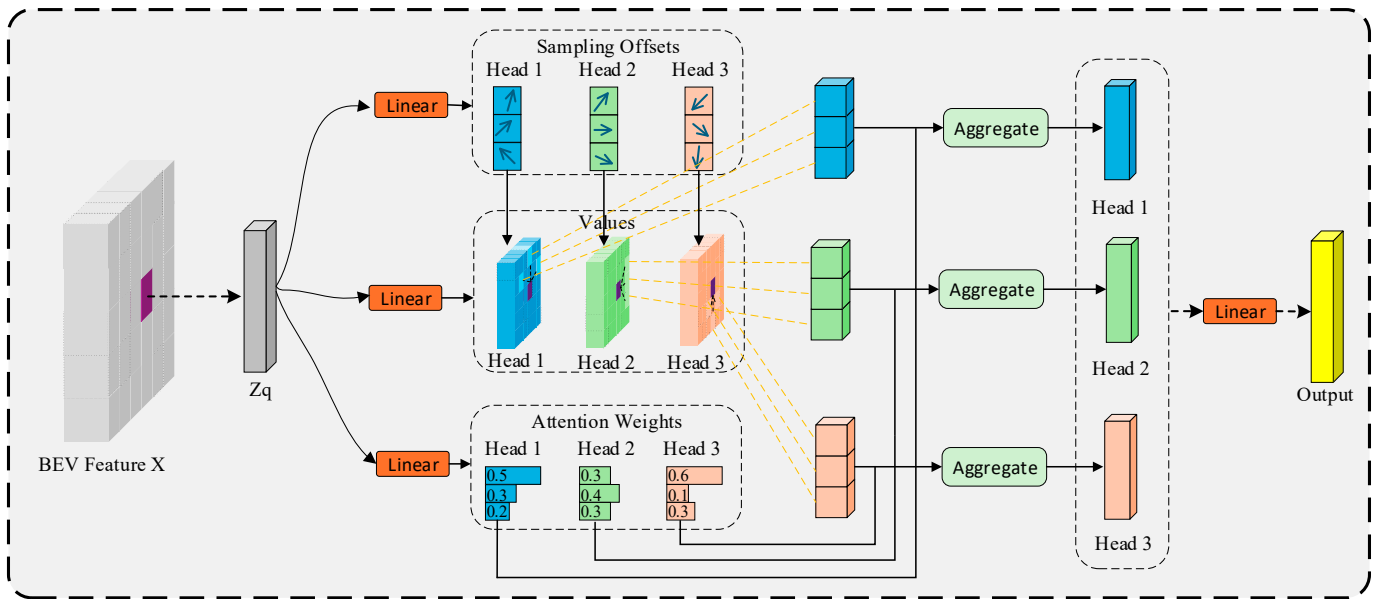


Figure 2. The proposed BEV Enhancement Module. The input is the BEV feature obtained from the view transformation module. We design three independent linear layers to respectively predict: (1) the attention weight of each head in the multi-head attention mechanism, (2) k groups of feature offsets for each head, and (3) the corresponding weights for the features acquired from these offsets. Finally, feature interaction and enhancement are performed based on the predicted result.

After processing through the Deformable Attention Module, we obtain the enhanced BEV feature F_{Be} , which emphasizes key localization elements—namely, road vector elements. For the above content, we have provided the pseudocode, as shown in Algorithm 1.

Algorithm 1: BEV Enhancement Module

Input:

$F_B \in \mathbb{R}^{H \times W \times C}$: BEV feature

$p_q \in \mathbb{R}^{H \times W \times 2}$: A set of reference points

K & M : The number of sampling points and the number of attention heads

W_q : Along with a randomly initialized, differentiable parameter matrix.

Output:

$F_{Be} \in \mathbb{R}^{H \times W \times C}$: The BEV features output by the BEV Enhancement Module

1: $q \leftarrow F_B \otimes W_q$

2: $V_m F_B$ & Δp_{mqk} & $W_{mqk} \leftarrow MLP_1(q) \text{ \& } MLP_2(q) \text{ \& } MLP_3(q)$

3: $head_1$ & $head_2$ & $head_3 \dots \leftarrow \sum_{k=1}^K (W_{mqk} \cdot V_m F_B(p_q + \Delta p_{mqk}))$

4: $F_{Be} \in \mathbb{R}^{H \times W \times C} \leftarrow MLP(Contact(head_1, head_2, head_3 \dots))$

3.3. Navigation Map Encoder

In navigation maps, dense road elements form the primary structural components. These elements can be regarded as high-frequency detail information due to their sharp boundaries, fine-grained edges, and structured geometric patterns. Unlike the BEV Feature Enhancement Module described in the previous section—which emphasizes high-level semantic representations of road features through supervisory constraints—the road elements in surround-view camera images do not inherently correspond to high-frequency representations. In navigation maps, however, accurately capturing such high-frequency structural details is essential. This necessity arises for two main reasons. First, the BEV features obtained in the previous stage already encode rich road semantics. For reliable cross-modal

registration, these semantic features must be precisely aligned with the fine-grained structural details present in navigation maps. Second, detailed edge and structural cues play a critical role in cross-modal feature fusion. Losing such high-frequency information would weaken alignment accuracy and degrade localization performance. To preserve detailed structural information while maintaining efficient feature transformation, we introduce an Invertible Neural Network (INN) module [54]. The key advantage of INNs lies in their bidirectional mapping property, which enables information-preserving transformations between input and output representations. By enforcing invertibility, the INN significantly reduces information loss during feature extraction and ensures more faithful preservation of fine-grained structural details. Moreover, since the INN is implemented using convolutional neural network (CNN)-based affine coupling layers, it maintains high computational efficiency. This characteristic makes it particularly suitable for real-time localization systems, where both accuracy and efficiency are critical. Accordingly, we adopt an INN block composed of affine coupling layers. In each invertible layer, the transformation is defined as follows:

$$\begin{aligned}\Phi_{l,k+1}^S[c+1:C] &= \Phi_{l,k}^S[c+1,C] + \Gamma_1(\Phi_{l,k}^S[1:c]), \Phi_{l,k+1}^S[1:C] = \Phi_{l,k}^S[1:C] \odot \exp \\ &(\Gamma_2(\Phi_{l,k+1}^S[c+1:C])) + \Gamma_3(\Phi_{l,k+1}^S[c+1:C]), \Phi_{l,k+1}^S = \mathcal{CAT}\{\Phi_{l,k+1}^S[1:C], \Phi_{l,k+1}^S[c+1:C]\}\end{aligned}\quad (4)$$

Here, $\odot$ denotes the Hadamard product, $\Phi_{l,k}^K[1:c] \in R^{H \times W \times C}$ denotes the input features of the first to the c -th channels of the k ($k = 1, \dots, K$)-th reversible layer. $\mathcal{CAT}(\cdot)$ represents the channel concatenation operation, and Γ_i ($i = 1, \dots, 3$) represents an arbitrary mapping function. Considering the trade-off between computational efficiency and feature extraction capability, we adopt the bottleneck residual block (BRB) from MobileNetV2 as Γ_i . For the above content, we have provided the pseudocode, as shown in Algorithm 2.

Algorithm 2: INN block

Input:

$X \in R^{C \times H \times W}$: Navigation map features encoded by a CNN

Output:

$Y \in R^{C \times H \times W}$: Navigation map features encoded by the INN

Γ_i : Bottleneck Residual Block from MobileNetV2

$s(\cdot), t(\cdot)$: Scale and translation functions implemented as Γ_i

1: $X1 \leftarrow X[0:C1, :, :], X2 \leftarrow X[C1:C, :, :]$

2: $Y1 \leftarrow X1, Y2 \leftarrow X2 \odot \exp(s(X1)) + t(X1)$

3: $Y \leftarrow \text{concat}(Y1, Y2) \in R^{C \times H \times W}$

3.4. Neural Localization Module

This section describes how cross-modal feature registration between BEV features and navigation map features is achieved. Recent studies [14,15] typically concatenate features from different modalities to obtain a unified representation. However, such direct fusion strategies overlook modality discrepancies and fail to explicitly model global spatial correlations across modalities. To overcome these limitations, we propose a BMF module that adaptively enhances the interaction between BEV features and navigation map features from a global perspective. Specifically, our objective is to achieve precise alignment of corresponding road elements across the two modalities, thereby improving cross-modal consistency and registration accuracy. The proposed cross-modal fusion framework consists of two key components: Feature-Adaptive Alignment and Feature-Adaptive Weighted Fusion. As shown in Figure 3, we first use a 1×1 convolution to

project both feature maps into an appropriate channel dimension and concatenate them. The fused features are then passed into two parallel branches. The first branch performs the feature alignment operation, which aims to identify and align the road information in both BEV and navigation map features. The alignment process first generates a 2D optical flow field ψ , composed of pixel-wise coordinate offsets. This flow field ψ is derived from the spatial discrepancies and correlation-based displacements between F_{be} and F_{Φ} . Using the computed flow field ψ , the two features are spatially aligned, which can be expressed as:

$$F_{Be} = \text{Warp}(F_{be}, \psi) \quad (5)$$

where $\text{Warp}(\cdot)$ denotes the bilinear interpolation warping function, which applies the calculated offsets to ensure spatial consistency between the features. The second branch of our cross-modal fusion framework is the Feature-Adaptive Weighted Fusion module. Given F_{be} and F_{Φ} , the concatenated feature F_a is first passed through 1×1 convolution to reduce its channel dimension to 1, followed by a sigmoid activation to produce the attention weight map W . This process can be expressed as:

$$W = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\text{Concat}(F_{be}, F_{\Phi}))) \quad (6)$$

where $\text{Sigmoid} = 1/e^{-x} + 1$. Finally, the fused feature is obtained by weighting the input features according to W .

$$F_{loca} = W \odot F_{Be} + (1 - W) \odot F_{\Phi} \quad (7)$$

For the above content, we have provided the pseudocode, as shown in Algorithm 3.

Algorithm 3: BMF

Input:

$F_{Be}, F_{\Phi} \in R^{H \times W \times C}$: BEV feature and Navigation map features

Output:

$F_{loca} \in R^{C \times H \times W}$

1: $\psi \in R^{H \times W} \leftarrow \text{Conv}(\text{Contact}(F_{Be}, F_{\Phi}))$

2: $F_{Be} \leftarrow F_{Be} + \psi$

3: $W \leftarrow \text{Sigmoid}(\text{Conv}_{1 \times 1}(\text{Concat}(F_{be}, F_{\Phi})))$

4: $F_{loca} \leftarrow W \odot F_{Be} + (1 - W) \odot F_{\Phi}$

3.5. Loss Function

In obtaining the fused features, we derive the global vehicle pose from them. A noisy prior initial pose $\hat{p} = (x, y, \theta)$ is given, and our goal is to estimate the optimal pose variation $\hat{\zeta} = (\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{\theta})$ the 2D navigation map. Since the ground-truth pose p is available, we apply a random perturbation of approximately ten meters to each ground-truth pose to simulate noisy initial poses provided by low-cost sensors such as GPS and IMU. Given the ground-truth pose variation $\zeta = (\Delta x, \Delta y, \Delta\theta)$ and the estimated value $\hat{\zeta} = (\Delta\hat{x}, \Delta\hat{y}, \Delta\hat{\theta})$, the pose loss function is defined as follows:

$$\mathcal{L}_l = \|\hat{p} + \hat{\zeta}, p\|_{S1} = \|\hat{p} + \hat{\zeta}, \hat{p} + \zeta\|_{S1} = \|\hat{\zeta}, \zeta\|_{S1} = \gamma(\|\Delta\hat{x}, \Delta x\|_{S1} + \|\Delta\hat{y}, \Delta y\|_{S1}) + \|\Delta\hat{\theta}, \Delta\theta\|_{S1} \quad (8)$$

In the above formulation, γ represents the balancing weight between the position loss and the rotation loss, where $\|\cdot\|_{S1}$ denotes the L2 Euclidean norm. We estimate the pose variation rather than directly predicting the absolute global pose because, over long distances, large numerical discrepancies between global poses inevitably arise. Such

excessive differences reduce the number of learnable samples and may lead to inaccurate localization. Detailed experimental results validating this design choice will be provided in the Experiment section.

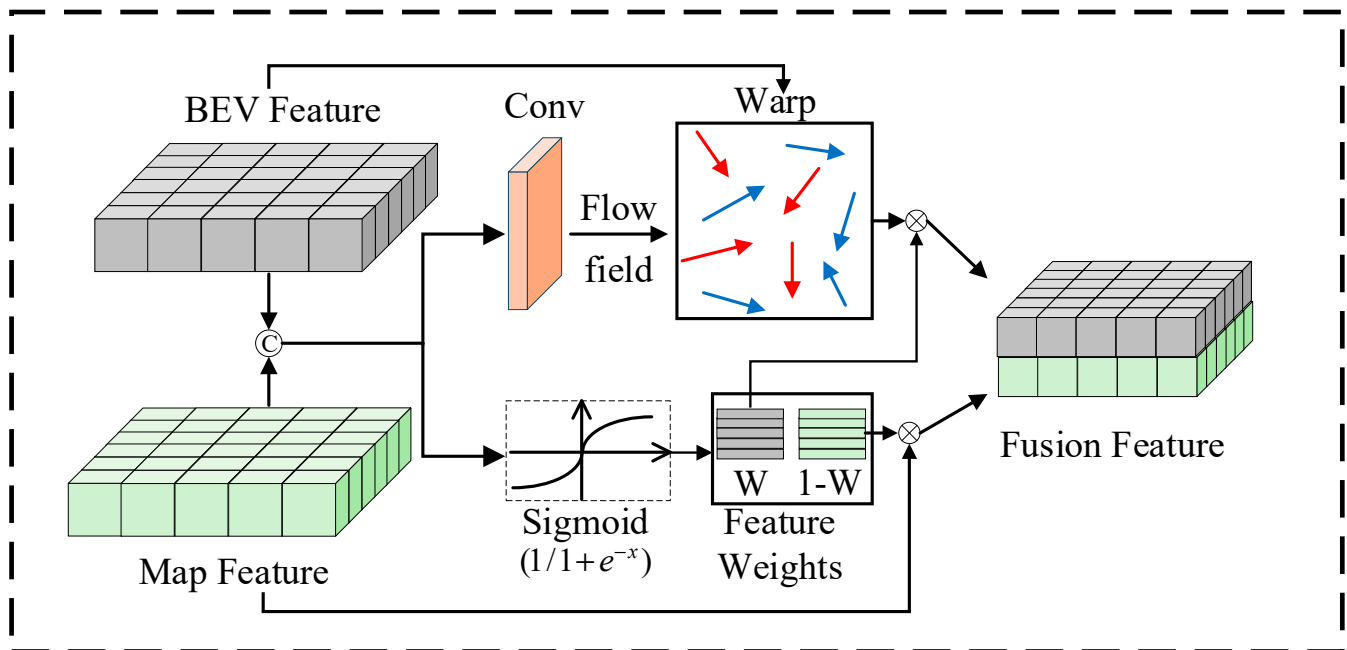


Figure 3. The proposed BMF Module. The inputs are the BEV features and the SD map features. We first perform a concatenation operation on the two feature sets, followed by offset prediction and a sigmoid activation to obtain the feature offsets and feature weights. Finally, the feature alignment process is completed based on the predicted results.

To better supervise BEV feature learning of road elements, we introduce the online end-to-end vector map construction task (MapTR [12]) as an auxiliary supervisory constraint in Section B. Following its hierarchical bipartite matching strategy, let $\hat{Y} = \{\hat{q}_i\}_{i=1}^N$ denote the set of predicted map elements, and $Y = \{q_i\}_{i=1}^N$ denote the set of ground-truth map elements. We aim to find the optimal instance-level matching between the two sets:

$$\hat{\pi} = \arg \min_{\pi \in \Pi_N} \sum_{i=1}^N \mathcal{L}_{nq}(\hat{q}_{\pi(i)}, q_i) \quad (9)$$

where $\pi \in \Pi_N$ represents a permutation of the points in a map instance, and $\mathcal{L}_{nq}(\hat{q}_{\pi(i)}, q_i)$ denotes the matching cost between the predicted and ground-truth map elements. The optimal assignment of instances and points is then determined using the Hungarian algorithm [55]. For the online vector map construction task, the total loss is composed of three components: instance loss, point loss, and point-direction loss, expressed as:

$$\mathcal{L}_m = \beta_c \cdot \mathcal{L}_{class} + \beta_p \cdot \mathcal{L}_{point} + \beta_d \cdot \mathcal{L}_{dense} \quad (10)$$

where β_c , β_p and β_d are weights balancing the different loss terms. Accordingly, our overall training objective is defined as:

$$\mathcal{L} = \lambda_c \cdot \mathcal{L}_{class} + \lambda_p \cdot \mathcal{L}_{point} + \lambda_d \cdot \mathcal{L}_{dense} + \lambda_l \cdot \mathcal{L}_l \quad (11)$$

where λ_c , λ_p , λ_d and λ_l are adaptive learning weights balancing the contributions of each component.

4. Results

4.1. Experimental Settings

4.1.1. Datasets

To comprehensively evaluate the performance of NMLoNet, we conducted experiments on the nuScenes [56] and Argoverse [57] datasets. The nuScenes dataset was collected in Boston and Singapore, where each frame's ground-truth vehicle pose was computed using a Monte Carlo localization scheme that integrates LiDAR point clouds and odometry measurements within an offline HD LiDAR map [4]. The dataset contains 850 sequences comprising approximately 34k frames in total. We follow the official split configuration, using 150 sequences (around 8k frames) for validation, which serves as our evaluation benchmark. The Argoverse dataset consists of 113 scenes recorded in Miami and Pittsburgh, among which 65 scenes are used for training and 24 scenes for validation and performance evaluation. To obtain the corresponding navigation map data, we extract geographic map information from OpenStreetMap (OSM) and rasterize it into a grid representation. To ensure precise alignment between each image frame and its associated navigation map, we adopt the BLOS-BEV [58] map-matching strategy. Specifically, we first convert map coordinates from EPSG:4326 [59] to EPSG:3857 [59] and then back to EPSG:4326 to obtain correctly aligned latitude–longitude coordinates, followed by appropriate scaling operations for spatial consistency. Furthermore, to simulate a noisy initial pose $\hat{p}$ for each frame, consistent with the procedure described in [15], we extract a map patch centered on the ego-vehicle from the rasterized navigation map. We then apply random rotation and translation perturbations $t \in [-15 \text{ m}, 15 \text{ m}]$ within the noise range typically observed in low-cost sensors such as GPS and IMU. Finally, the central region of the perturbed patch is cropped and used as the navigation map input to our network.

4.1.2. Implementation Details

Our proposed NMLoNet utilizes six surround-view images as visual inputs, each with an image resolution of 128×352 . The perception range of the ego vehicle in the BEV (Bird's-Eye View) space is defined as $[-50 \text{ m}, 50 \text{ m}]$ meters along the longitudinal axis and $[-15 \text{ m}, 15 \text{ m}]$ meters along the lateral axis, with a grid resolution of 0.5 m. The depth perception range is set to $[4 \text{ m}, 60 \text{ m}]$ meters, ensuring sufficient coverage for both near-field and far-field regions in autonomous driving scenarios. For the extent of the navigation map, this paper selects twice the perception range $[-100 \text{ m}, 100 \text{ m}]$ and $[-30 \text{ m}, 30 \text{ m}]$.

All experiments are conducted using eight NVIDIA Tesla V100 GPUs. The proposed model is optimized using the AdamW optimizer, with a weight decay of 1×10^{-4} , a batch size of 8, and an initial learning rate of 1×10^{-5} . On the nuScenes dataset, the model is trained for 110 epochs until convergence, while on the Argoverse dataset, 30 epochs are sufficient for stable performance.

4.2. Comparisons with State-of-the-Art Methods

4.2.1. Results on nuScenes

As shown in Table 1, our method consistently outperforms all competing approaches in terms of both positional and orientation accuracy. In particular, for localization accuracy within 1 m, NMLoNet achieves 24.9%, exceeding MapLocNet [15] by 4.9 percentage points. This improvement demonstrates that our model effectively captures fine-grained roadway semantics in BEV space as well as detailed structural information from navigation maps. These features provide stable and environment-invariant cues for localization, thereby substantially reducing overall positional error. Notably, under the monocular camera setting, our method still achieves an accuracy of 11.3% within 1 m, outperforming all baseline approaches. Although monocular images inherently produce weaker BEV representations

compared to surround-view systems due to limited spatial coverage, NMLoNet is still able to guide BEV features to emphasize road semantic elements. This facilitates precise cross-modal alignment with 2D navigation map features and ultimately enables accurate localization performance.

Table 1. Localization results on the nuScenes dataset on eight NVIDIA V100 GPUs.

Approach	Cams	Position Recall (m)			Orientation Recall (°)			Param (M)	GFLOPs	FPS
		1	2	5	1	2	5			
OrienterNet	1	5.8	18.9	52.8	32.4	41.5	65.6	54.9	161.9	8.1
MapLocNet	1	7.14	22.39	60.4	35.1	60.4	87.5	20.7	49.81	23.8
NMLoNet	1	11.3	27.6	63.2	37.2	68.6	91.8	23.8	53.45	24.6
U-BEV	6	16.8	41.6	71.3						
MapLocNet	6	20.1	45.54	77.7	58.6	84.1	96.2	22.9	56.15	18.2
NMLoNet	6	24.9	58.3	85.1	65.1	86.2	96.7	23.8	55.32	22.7

Figure 4 illustrates the temporal distribution of localization errors produced by our method, including both lateral and longitudinal components. For lateral localization, NMLoNet maintains deviations within approximately 30 cm. This high level of accuracy can be attributed to the strong geometric constraints provided by road semantics, such as lane boundaries and road edges, which offer reliable lateral alignment cues during cross-modal registration. In contrast, longitudinal errors remain at the decimeter level. This is primarily because fewer distinctive semantic landmarks—such as crosswalks or traffic poles—are available to constrain alignment along the longitudinal direction. As a result, residual errors in this direction are relatively larger. Nevertheless, in practical autonomous driving systems, downstream modules such as motion control and path planning generally impose less stringent requirements on longitudinal precision compared to lateral accuracy. Consequently, the practical impact of this limitation is mitigated.

4.2.2. Results on Argoverse

To further validate the effectiveness of the proposed model, we conduct additional experiments on the Argoverse dataset. The model is trained using the same hyperparameter settings as those adopted for nuScenes. For fairness, all comparative methods are retrained under identical experimental conditions.

As reported in Table 2, our method achieves the best performance under both monocular and surround-view configurations. Specifically, for localization accuracy within 1 m, NMLoNet attains 27.4%, outperforming MapLocNet [15] by 4.1 percentage points. Furthermore, the trajectory error curves shown in Figure 5 demonstrate that our approach consistently produces stable and accurate localization estimates over time. These results collectively confirm the robustness and strong generalization capability of our model across different datasets and sensor configurations.

Table 2. Localization results on the Argoverse dataset on eight NVIDIA V100 GPUs.

Approach	Cams	Position Recall (m)			Orientation Recall (°)			Param (M)	GFLOPs	FPS
		1	2	5	1	2	5			
OrienterNet	1	8.56	21.2	54.9	18.7	31.1	65.6	54.9	161.9	8.1
MapLocNet	1	9.12	27.6	66.3	41.2	66.5	90.3	20.7	49.81	23.8
NMLoNet	1	12.8	29.6	65.2	47.2	72.6	94.8	23.8	53.45	24.6
MapLocNet	6	23.3	47.2	79.3	62.4	86.3	96.24	22.9	56.15	18.2
NMLoNet	6	27.4	53.3	86.4	67.6	87.5	98.6	23.8	55.32	22.7

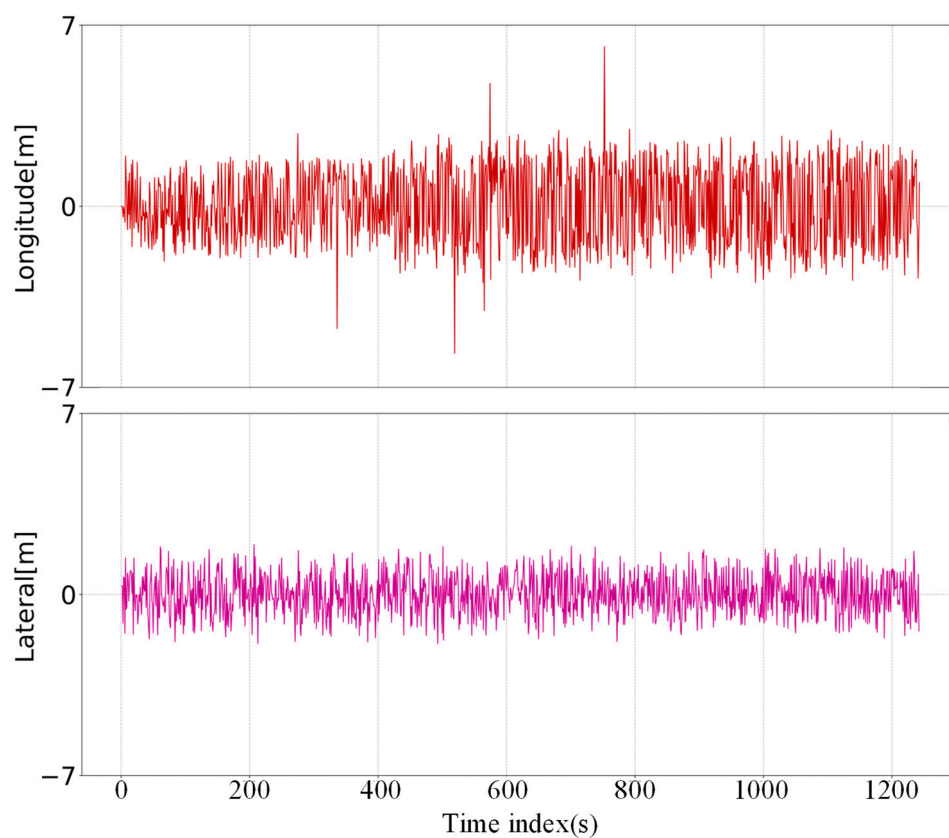


Figure 4. Model error on the nuScenes dataset.

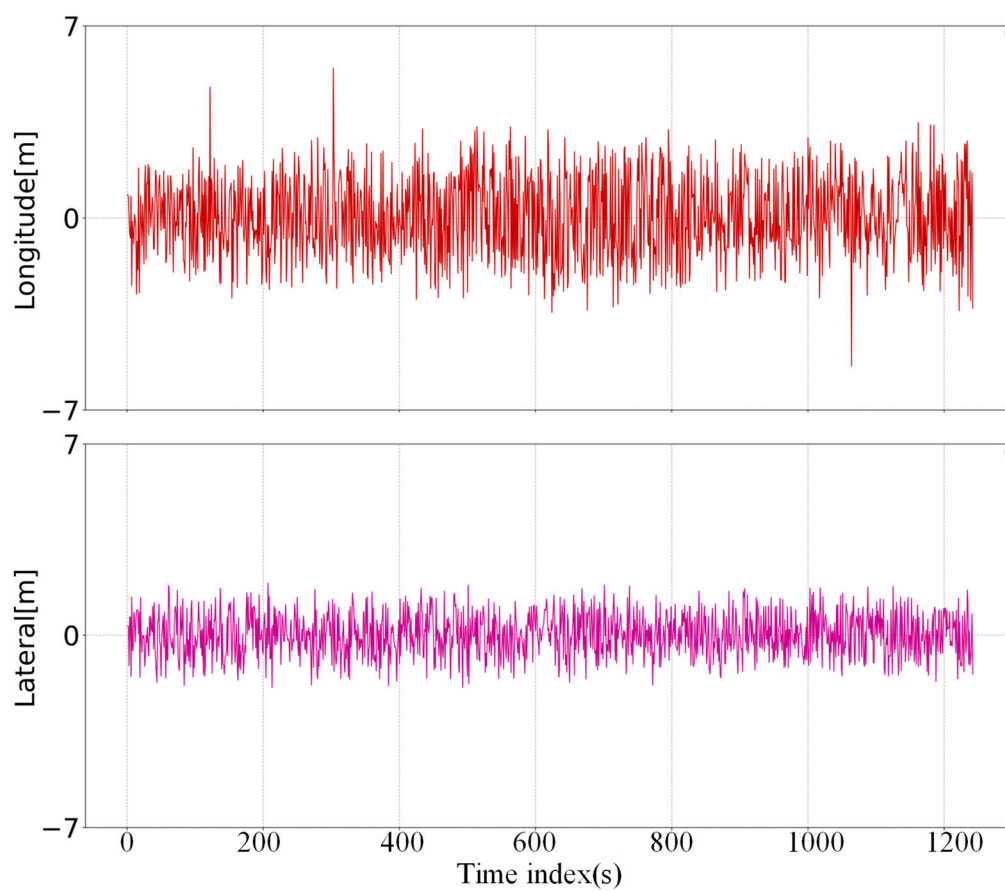


Figure 5. Model error on the Argoverse dataset.

4.2.3. BEV Visualization on the nuScene

The key objective of our proposed framework is to reduce the modality gap between BEV features and navigation map representations by encouraging BEV features to focus on static road elements in the environment. To validate this, we visualize the BEV feature heatmaps on the nuScenes dataset, as illustrated in Figure 6. The visualization results reveal that the extracted BEV features indeed concentrate on static roadway structures, such as lanes and boundaries, thereby confirming the rationality and effectiveness of the proposed NMLoNet.

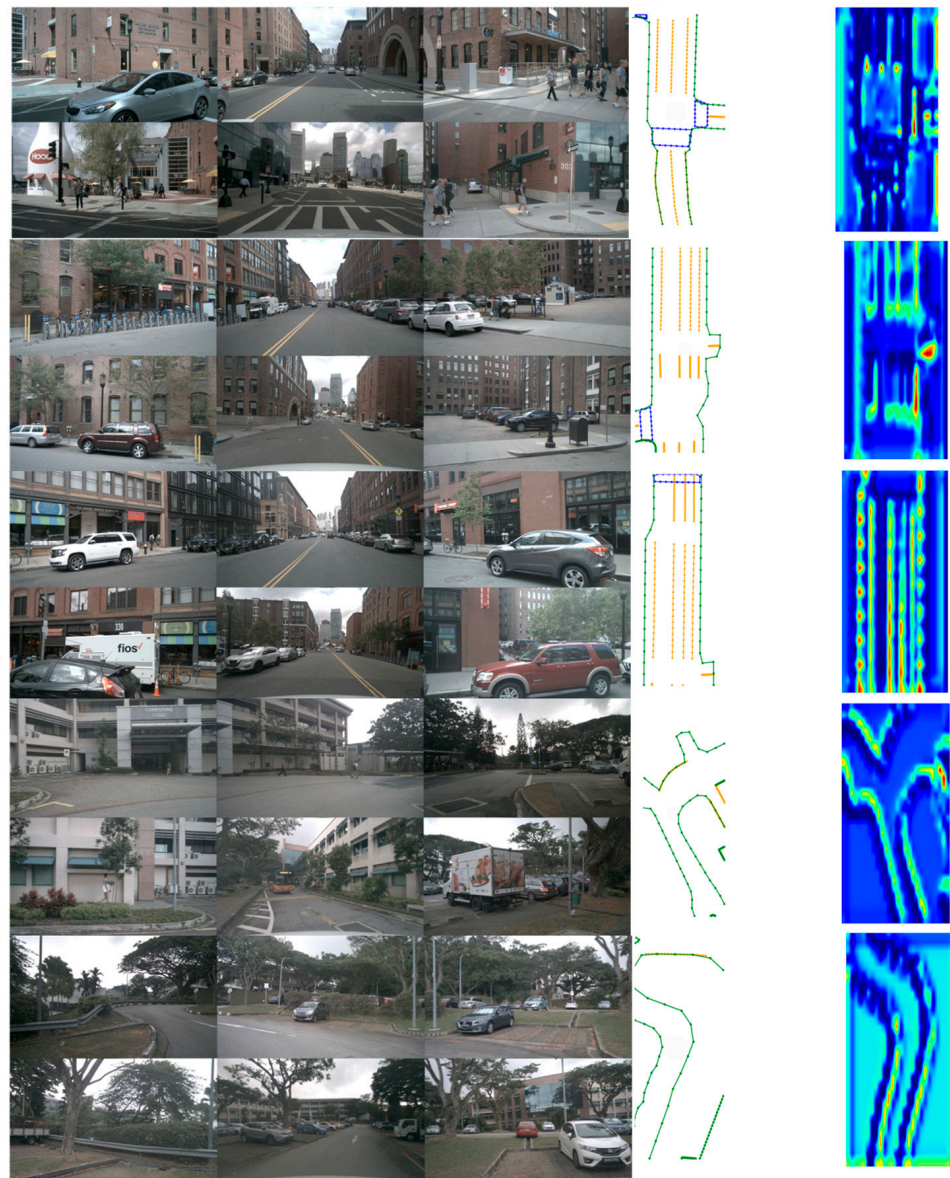


Figure 6. BEV feature heatmaps of our model on the nuScenes dataset. The middle image shows the real-time vector map output from the vector map module, and the right image shows the BEV feature heatmap.

4.3. Ablation Studies

This section presents extensive ablation studies to investigate the effectiveness of each component within NMLoNet as well as the influence of their corresponding hyperparameter settings. All experiments are conducted on the nuScenes dataset.

4.3.1. Effectiveness of the BEV Enhancement Module and INN

We conduct ablation studies on the nuScenes dataset to evaluate the contribution of each proposed module, and the results are summarized in Table 3. For the proposed BEV Feature Enhancement Module, the proportion of samples with a position error smaller than 1 m increases by 5%. This improvement demonstrates that the self-attention mechanism effectively refines the attention distribution within BEV features, guiding the model to focus on more informative and discriminative regions. An additional interpretable explanation is that, in scenarios with sparse lane markings, the module can adaptively emphasize critical semantic cues, thereby facilitating more robust feature learning and inference. Moreover, we observe that the BEV Feature Enhancement Module yields greater improvements in longitudinal accuracy than in lateral accuracy. This phenomenon can be attributed to the relatively limited availability of road semantic elements that provide longitudinal constraints, which makes it challenging for conventional models to capture these sparse cues. In contrast, the Deformable Attention mechanism incorporated in our module can dynamically attend to such informative regions through sparse, learnable sampling offsets, thereby enhancing localization performance—particularly along the longitudinal direction. For the INN module integrated into the Navigation Map Encoder, the model achieves a 4.6% improvement in the proportion of samples with position error below 1 m. The motivation for introducing the INN module is to better capture high-frequency structural information from navigation map images. Specifically, it enables the extraction of fine-grained road semantic details, which helps reduce the modality gap between navigation maps and visual BEV features. The ablation results further show that removing the INN module leads to a noticeable decline in localization accuracy. This confirms that the INN module enhances the model's ability to exploit detailed road semantic information, effectively narrowing the cross-modal discrepancy and improving overall positioning performance.

Table 3. Ablation study results on the nuScenes dataset evaluating the effectiveness of the BEV Enhancement Module and the INN for localization.

Model	Position Recall (m)			Orientation Recall (°)		
	1	2	5	1	2	5
BEV Enhancement	19.2	52.1	71.2	54	77.2	88.1
INN	19.6	56.8	72.9	56.2	77.8	90.3
Base	24.2	58.3	85.1	65.1	86.2	96.7

4.3.2. Effectiveness of the Neural Localization Module

To evaluate the effectiveness of the proposed cross-modal feature fusion module, we design three comparative ablation settings: (1) direct feature concatenation, (2) optical-flow-based feature alignment only, and (3) adaptive feature weighting only. As reported in Table 4, directly concatenating features results in a 6-percentage-point decrease in localization accuracy. This significant performance drop indicates that naive feature fusion introduces representation misalignment across modalities, leading to increased localization errors. The result highlights the necessity of explicitly modeling cross-modal discrepancies rather than relying on simple feature aggregation. For the other two variants, the experimental results show that removing either the feature alignment submodule or the adaptive weighting submodule leads to noticeable performance degradation. This observation confirms that both components are critical for effective cross-modal feature fusion. Specifically, the alignment module enforces spatial consistency between BEV and navigation map features, while the adaptive weighting module enables balanced, context-aware information integration by dynamically modulating the contribution of each modality.

Table 4. Ablation study results on the nuScenes dataset evaluating the effectiveness of the Neural Localization Module.

Model	Position Recall (m)			Orientation Recall (°)		
	1	2	5	1	2	5
Direct Feature Fusion	18.2	53.8	79.3	34.2	76.4	90.2
Optical Flow Feature Alignment	21.9	55.3	81.2	35.8	75.8	91.4
Adaptive Feature Weighting Module	22.5	56.5	82.2	35.7	77.6	91.3
Base	24.2	58.3	85.1	65.1	86.2	96.7

4.3.3. Effectiveness of the Different Road Elements

Standard navigation maps typically lack explicit road boundary annotations. However, road boundaries provide critical semantic cues that play a key role in the auxiliary supervision of BEV feature learning. To investigate their influence on localization performance, we conduct ablation studies by explicitly incorporating road boundary information into the map representation. As shown in Table 5, the inclusion of road boundaries yields a measurable improvement in localization accuracy. This gain stems from the fact that road boundaries constitute an indispensable component of static semantic structure: they not only serve as geometric anchors for distinguishing between opposing and adjacent lanes but also enable more robust modeling of scene-level static semantics, thereby enhancing cross-modal alignment fidelity.

Table 5. Ablation study results on the nuScenes dataset evaluating the effectiveness of the different road elements.

Model	Position Recall (m)			Orientation Recall (°)		
	1	2	5	1	2	5
No Road Boundary	23.8	56.7	86.1	60.2	84.7	97.2
All Elements	24.2	58.3	85.1	65.1	86.2	96.7

5. Discussion

Experimental results demonstrate that NMLoNet achieves a favorable trade-off between computational efficiency and localization accuracy within an end-to-end vehicle localization framework. Specifically, the proposed method improves localization accuracy by approximately 11% under monocular settings and by 24% under surround-view configurations. The integration of online map constraints together with the BEV Feature Enhancement Module significantly strengthens the representation of structural environmental cues in bird's-eye-view space, enabling more accurate encoding of localization-relevant road elements. In addition, the Invertible Neural Network (INN) module and the Bidirectional Modality Fusion (BMF) module effectively reduce the modality gap between BEV features and navigation map representations, thereby improving cross-modal alignment and overall localization performance. Nevertheless, several limitations should be acknowledged. First, the current framework relies exclusively on environmental perception (i.e., BEV features derived from visual input) and sparse navigation maps, without incorporating vehicle motion cues such as IMU measurements or wheel encoder data. These motion-related signals could provide complementary temporal constraints and further enhance localization robustness, particularly in visually ambiguous or dynamic scenarios. Second, due to the extreme sparsity of publicly available navigation maps (e.g., OpenStreetMap), our method depends more heavily on real-time perceptual features than on map-specific structural details. As a result, its localization accuracy does not yet match that of approaches relying on high-definition (HD) maps, which contain dense geometric and

semantic annotations. However, this characteristic also highlights a key advantage of our approach: it does not strongly depend on rich map priors. Because standard navigation maps contain substantially less semantic and geometric information than HD maps, our framework is inherently more adaptable to map-scarce or resource-constrained environments. While map sparsity limits the achievable map-driven accuracy, it simultaneously enhances scalability and generalizability across large-scale deployment scenarios.

In future work, we plan to incorporate additional sensor modalities—such as IMU, GPS, and wheel odometry—into the model input to provide richer motion-aware contextual information. By integrating perceptual and dynamic constraints within a unified framework, we aim to further improve localization robustness and accuracy under diverse real-world conditions.

6. Conclusions

In conclusion, we have presented NMLoNet, an end-to-end neural localization framework that achieves high-precision vehicle localization using only six-view surround images and standard 2D navigation maps—eliminating the need for expensive, pre-built HD maps. By explicitly addressing the modality gap between BEV visual features and map geometry through deformable attention, invertible detail extraction, and adaptive cross-modal fusion, our method delivers both accuracy and robustness under real-world driving conditions. The effectiveness of NMLoNet is rigorously validated on two large-scale autonomous driving benchmarks: nuScenes and Argoverse. Under monocular settings, our approach reduces the average localization error by 11% compared to the best prior map-based method; under full surround-view configurations, this improvement increases to 24%, achieving a median position error of 0.38 m on nuScenes—comparable to HD-map-dependent systems. Moreover, in challenging scenarios such as dense urban canyons, nighttime conditions, and highly dynamic traffic (where over 30 vehicles are present per frame), NMLoNet maintains stable performance with less than 5% degradation in success rate, demonstrating its practical reliability. These results confirm that NMLoNet not only advances the state of the art in map-efficient localization but also provides a scalable and cost-effective solution for real-world deployment of intelligent vehicles. Future work will explore integration with sparse LiDAR or GPS priors to further enhance robustness in extreme environments such as long tunnels or heavy occlusions.

Author Contributions: Conceptualization, Q.Y. and Y.L.; Software, Q.Y.; Validation, Q.Y.; Formal analysis, Y.L.; Investigation, Y.L.; Resources, Q.Y.; Writing—review and editing, Q.Y. and Y.L.; Supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by National Natural Science Foundation of China (52272418, U22A20100).

Data Availability Statement: The nuScenes dataset used in this study is publicly available at <https://www.nuscenes.org/> (accessed 26 February 2025). The argoverse dataset is publicly available at <https://www.argoverse.org/> (accessed 26 February 2025). The code for this project will be made available upon publication.

Acknowledgments: The authors would like to acknowledge the contributions of the open-source community for the software packages used in this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Han, J.; Ju, Z.; Chen, X.; Yang, M.; Zhang, H.; Huai, R. Secure Operations of Connected and Autonomous Vehicles. *IEEE Trans. Intell. Veh.* **2023**, *8*, 4484–4497. [[CrossRef](#)]

2. Claussmann, L.; Revilloud, M.; Gruyer, D.; Glaser, S. A review of motion planning for highway autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1826–1848. [\[CrossRef\]](#)
3. Otegui, J.; Bahillo, A.; Lopetegi, I.; Díez, L.E. Evaluation of experimental GNSS and 10-DOF MEMS IMU measurements for train positioning. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 269–279. [\[CrossRef\]](#)
4. Qin, T.; Li, P.; Shen, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [\[CrossRef\]](#)
5. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [\[CrossRef\]](#)
6. Shan, T.; Englot, B. LeGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 4758–4765.
7. Shan, T.; Englot, B.; Meyers, D.; Wang, W.; Ratti, C.; Rus, D. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 24–29 October 2020; pp. 5135–5142.
8. Li, Y.; Hu, Z.; Cai, Y.; Wu, H.; Li, Z.; Sotelo, M.A. Visual map-based localization for intelligent vehicles from multi-view site matching. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1068–1079. [\[CrossRef\]](#)
9. Cai, Y.; Lu, Z.; Wang, H.; Chen, L.; Li, Y. A lightweight feature map creation method for intelligent vehicle localization in urban road environments. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 8503115. [\[CrossRef\]](#)
10. Li, Y.; Feng, F.; Cai, Y.; Li, Z.; Sotelo, M.A. Localization for intelligent vehicles in underground car parks based on semantic information. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 1317–1332. [\[CrossRef\]](#)
11. Li, Q.; Wang, Y.; Wang, Y.; Zhao, H. HDMapNet: An online HD map construction and evaluation framework. In Proceedings of the IEEE International Conference on Robotics and Automation, Philadelphia, PA, USA, 23–27 May 2022; pp. 4628–4634.
12. Liao, B.; Chen, S.; Zhang, Y.; Jiang, B.; Cheng, T.; Zhang, Q.; Liu, W.; Huang, C.; Wang, X. MAPTR: Structured modeling and learning for online vectorized HD map construction. *arXiv* **2023**, arXiv:2208.14437. [\[CrossRef\]](#)
13. Sarlin, P.-E.; Trulls, E.; Pollefeys, M.; Hosang, J.; Lynen, S. SNAP: Self-supervised neural maps for visual positioning and semantic understanding. *arXiv* **2023**, arXiv:2306.05407. [\[CrossRef\]](#)
14. Sarlin, P.-E.; DeTone, D.; Yang, T.-Y.; Avetisyan, A.; Straub, J.; Malisiewicz, T.; Buló, S.R.; Newcombe, R.; Kotschieder, P.; Balntas, V. OrienterNet: Visual localization in 2D public maps with neural matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21632–21642.
15. Wu, H.; Zhang, Z.; Lin, S.; Mu, X.; Zhao, Q.; Yang, M.; Qin, T. MapLocNet: Coarse-to-fine feature registration for visual re-localization in navigation maps. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 13198–13205.
16. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [\[CrossRef\]](#)
17. Xu, W.; Zhang, F. FAST-LIO: A fast, robust lidar-inertial odometry package by tightly-coupled iterated Kalman filter. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3317–3324. [\[CrossRef\]](#)
18. Zhang, J.; Singh, S. Visual-lidar odometry and mapping: Low-drift, robust, and fast. In Proceedings of the IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 2174–2181.
19. Wang, Z.; Fang, J.; Dai, X.; Zhang, H.; Vlacic, L. Intelligent vehicle self-localization based on double-layer features and multilayer LIDAR. *IEEE Trans. Intell. Veh.* **2020**, *5*, 616–625. [\[CrossRef\]](#)
20. Li, Y.; Cai, Y.; Malekian, R.; Wang, H.; Sotelo, M.A.; Li, Z. Creating navigation map in semi-open scenarios for intelligent vehicle localization using multi-sensor fusion. *Expert Syst. Appl.* **2021**, *184*, 115543. [\[CrossRef\]](#)
21. Qin, T.; Zheng, Y.; Chen, T.; Chen, Y.; Su, Q. A light-weight semantic map for visual localization towards autonomous driving. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 11248–11254.
22. Barbosa, B.H.G.; Bhatt, N.P.; Khajepour, A.; Hashemi, E. Adaptive and soft constrained vision-map vehicle localization using Gaussian processes and instance segmentation. *Expert Syst. Appl.* **2025**, *264*, 125790. [\[CrossRef\]](#)
23. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
24. Walch, F.; Hazirbas, C.; Leal-Taixé, L.; Sattler, T.; Hilsenbeck, S.; Cremers, D. Image-based localization using LSTMs for structured feature correlation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 627–637.
25. Brahmbhatt, S.; Gu, J.; Kim, K.; Hays, J.; Kautz, J. Geometry-aware learning of maps for camera localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2616–2625.

26. Zhu, Y.; Gao, R.; Huang, S.; Zhu, S.-C.; Wu, Y.N. Learning neural representation of camera pose with matrix representation of pose shift via view synthesis. *arXiv* **2021**, arXiv:2104.01508. [[CrossRef](#)]
27. Xue, F.; Wu, X.; Cai, S.; Wang, J. Learning multi-view camera relocation with graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11372–11381.
28. Lu, W.; Zhou, Y.; Wan, G.; Hou, S.; Song, S. L3-Net: Towards learning based LiDAR localization for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6382–6391.
29. Zhou, Y.; Wan, G.; Hou, S.; Yu, L.; Wang, G.; Rui, X.; Song, S. DA4AD: End-to-end deep attention-based visual localization for autonomous driving. *arXiv* **2020**, arXiv:2003.03026.
30. Ma, W.-C.; Tartavull, I.; Bârsan, I.A.; Wang, S.; Bai, M.; Mattyus, G.; Homayounfar, N.; Lakshmikanth, S.K.; Pokrovsky, A.; Urtasun, R. Exploiting sparse semantic HD maps for self-driving vehicle localization. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Macau, China, 4–8 November 2019; pp. 5304–5311.
31. Lindenberger, P.; Sarlin, P.-E.; Larsson, V.; Pollefeys, M. Pixel-perfect structure-from-motion with featuremetric refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 5967–5977.
32. Cattaneo, D.; Vaghi, M.; Ballardini, A.L.; Fontana, S.; Sorrenti, D.G.; Burgard, W. CMRNet: Camera to LiDAR-map registration. In Proceedings of the IEEE Intelligent Transportation Systems Conference, Auckland, New Zealand, 27–30 October 2019; pp. 1283–1289.
33. Miao, J.; Jiang, K.; Wang, Y.; Wen, T.; Xiao, Z.; Fu, Z.; Yang, M.; Liu, M.; Huang, J.; Zhong, Z.; et al. Poses as queries: End-to-end image-to-LiDAR map localization with transformers. *IEEE Robot. Autom. Lett.* **2024**, *9*, 803–810. [[CrossRef](#)]
34. Chang, M.-F.; Mangelson, J.; Kaess, M.; Lucey, S. HyperMap: Compressed 3D map for monocular camera registration. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 11739–11745.
35. Zhang, Z.; Xu, M.; Zhou, W.; Peng, T.; Li, L.; Poslad, S. BEV-Locator: An end-to-end visual semantic localization network using multi-view images. *arXiv* **2022**, arXiv:2211.14927. [[CrossRef](#)]
36. He, Y.; Liang, S.; Rui, X.; Cai, C.; Wan, G. EgoVM: Achieving precise ego-localization using lightweight vectorized maps. *arXiv* **2023**, arXiv:2307.08991.
37. Zhang, S.; Yang, T.; Chen, C. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5316–5325.
38. Shi, Y.; Li, H. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16989–16999.
39. Zhu, S.; Shah, M.; Chen, C. TransGeo: Transformer is all you need for cross-view image geo-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1152–1161.
40. Fervers, F.; Bullinger, S.; Bodensteiner, C.; Arens, M.; Stiefelwagen, R. Uncertainty-aware vision-based metric cross-view geolocalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21621–21631.
41. Panphattarasap, P.; Calway, A. Automated map reading: Image based localisation in 2-D maps using binary semantic descriptors. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 6341–6348.
42. Samano, N.; Zhou, M.; Calway, A. You are here: Geolocation by embedding maps and images. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12117, pp. 502–518.
43. Cho, Y.; Kim, G.; Lee, S.; Ryu, J.-H. OpenStreetMap-based LiDAR global localization in urban environment without a prior LiDAR map. *IEEE Robot. Autom. Lett.* **2022**, *7*, 4999–5006. [[CrossRef](#)]
44. Kwon, O.; Jung, D.; Kim, Y.; Ryu, S.; Yeon, S.; Oh, S.; Lee, D. WayIL: Image-based indoor localization with wayfinding maps. In Proceedings of the IEEE International Conference on Robotics and Automation, Yokohama, Japan, 13–17 May 2024; pp. 6274–6281.
45. Haklay, M.; Weber, P. OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [[CrossRef](#)]
46. Philion, J.; Fidler, S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12114, pp. 194–210.
47. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; Dai, J. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13698, pp. 1–18.
48. Chen, S.; Cheng, T.; Wang, X.; Meng, W.; Zhang, Q.; Liu, W. Efficient and robust 2D-to-BEV representation learning via geometry-guided kernel transformer. *arXiv* **2022**, arXiv:2206.04584.

49. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
50. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
51. Chu, F.; Li, H.; Xie, L.; Zhao, J. A survey of transformer architectures for autonomous driving. *Expert Syst. Appl.* **2026**, *299*, 130338. [[CrossRef](#)]
52. Haqiqi, M.M.E.; Arifin, A.S.; Satyawan, A.S. Enhancing Object Detection for Autonomous Vehicles in Low-Resolution Environments Using a Super-Resolution Transformer-Based Preprocessing Framework. *World Electr. Veh. J.* **2025**, *16*, 678. [[CrossRef](#)]
53. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2020.
54. Zhou, M.; Fu, X.; Huang, J.; Zhao, F.; Liu, A.; Wang, R. Effective pan-sharpening with transformer and invertible neural network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5502315. [[CrossRef](#)]
55. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [[CrossRef](#)]
56. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11618–11628.
57. Chang, M.-F.; Lambert, J.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; Lucey, S.; Ramanan, D.; et al. Argoverse: 3D tracking and forecasting with rich maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8740–8749.
58. Wu, H.; Zhang, Z.; Lin, S.; Qin, T.; Pan, J.; Zhao, Q.; Xu, C.; Yang, M. BLOS-BEV: Navigation map enhanced lane segmentation network, beyond line of sight. In Proceedings of the IEEE Intelligent Vehicles Symposium, Los Angeles, CA, USA, 9–12 June 2024; pp. 3212–3219.
59. International Association of Oil; Gas Producers (IOGP). *EPSG Geodetic Parameter Dataset, EPSG:4326—WGS 84 and EPSG:3857—WGS 84/Pseudo-Mercator*; IOGP: London, UK, 2024. Available online: <https://www.iogp.org/> (accessed on 26 February 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.