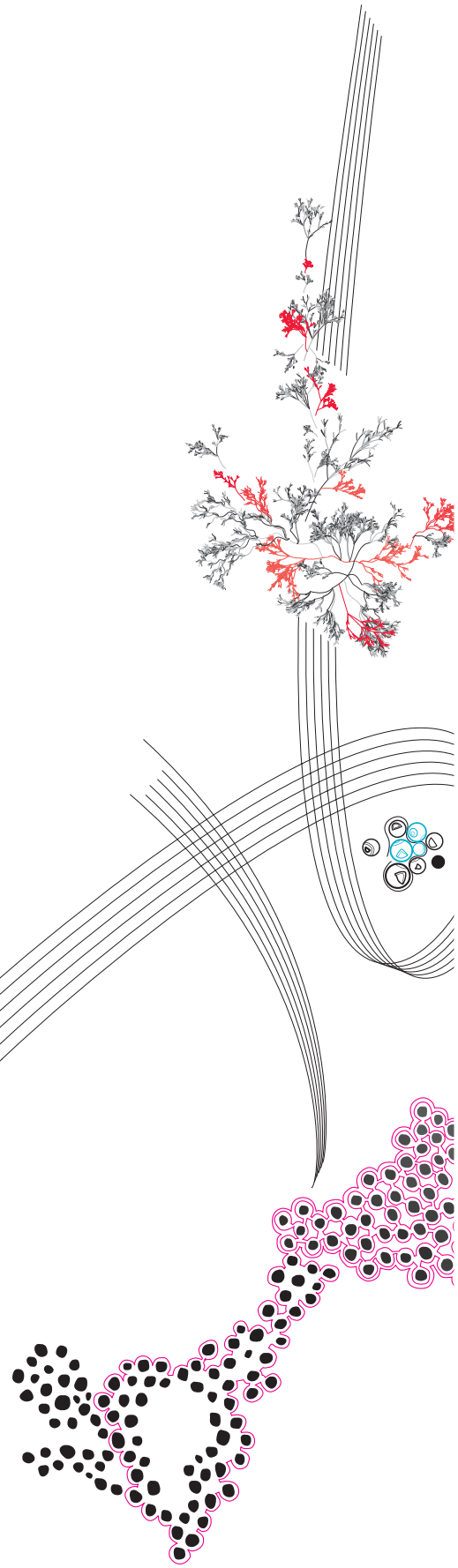

MSc Computer Science

Master Thesis

Exploring the Intersection of End-to-End HD Mapping and HD Map-Based Localization: A Survey with Implementation Perspectives

Ziyi Wang

Supervisor:
Maurice van Keulen (UT)
Lena Wild (Scania)
Rafael Valencia (Scania)

September, 2025

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

High-definition (HD) maps provide centimeter-level spatial accuracy essential for autonomous driving, yet traditional production workflows present significant scalability limitations due to resource-intensive data collection and manual processing. This thesis addresses the critical research gap by investigating the integration of two independently evolving domains: learning-based end-to-end HD mapping and HD map-based localization systems.

Current approaches develop along parallel trajectories, limiting system adaptability as real-world autonomous systems require both continuously updated maps and precise localization within evolving environments. Through comprehensive literature survey, experimental validation, and architectural framework design, this research demonstrates the fundamental interdependence between mapping and localization performance.

The comparative analysis reveals that both HD and standard-definition (SD) map-based approaches converge on transformer-based Bird's Eye View representations and cross-modal attention mechanisms, indicating that semantic-geometric correspondence learning represents the core technical challenge. HD map-based methods achieve centimeter-level accuracy under constrained conditions while SD map-based methods provide meter-level precision with superior tolerance to large pose uncertainties.

Experimental validation establishes the fundamental interdependence between mapping and localization through systematic pose perturbation analysis that simulates real-world GPS/IMU positioning uncertainties. Localization experiments demonstrate that training with realistic pose uncertainties significantly enhances performance stability compared to idealized scenarios, while HD mapping experiments reveal a dual-threshold degradation pattern where fine-scale perturbations induce gradual performance reduction and large-scale perturbations precipitate catastrophic collapse. These findings establish that accurate pose estimation enables superior mapping through effective historical map integration, whereas pose uncertainties fundamentally compromise spatial correspondence and prior information retrieval.

The proposed end-to-end architecture enables simultaneous HD mapping and localization through joint optimization, featuring hybrid prior map integration, dual-layer storage balancing stability with environmental responsiveness, and decoupled pose estimation maintaining computational efficiency.

This research provides the first comprehensive survey bridging HD mapping and localization domains, establishes quantitative performance benchmarks, and identifies optimal integration strategies. The findings inform next-generation autonomous driving systems capable of maintaining accuracy while adapting to dynamic operational environments.

Keywords: **High-definition maps, Autonomous Driving, Map-based Localization, HD Mapping, SLAM, BEV**

Chapter 1

Introduction

Autonomous driving systems are conventionally structured around three fundamental modules: **perception, planning, and control**. Among these, accurate vehicle localization serves as the critical foundation upon which all subsequent operational decisions depend [6]. Precise knowledge of vehicle pose within a global reference frame is essential for executing safe maneuvers, from basic lane-keeping to complex trajectory optimization in dynamic environments. Even localization errors at the decimeter level can potentially lead to unsafe vehicle behavior or collisions, particularly in dense urban settings characterized by intricate road geometries, dynamic traffic participants, and frequent visual occlusions [34].

While global navigation satellite systems (GNSS) provide position estimates with global coverage, their performance deteriorates significantly in challenging environments such as urban canyons, tunnels, and densely forested areas. In these scenarios, multipath effects and signal blockages produce positioning errors that substantially exceed the sub-10 cm accuracy threshold required for safe autonomous operation [34]. This fundamental limitation of GNSS technology necessitates complementary localization approaches for reliable autonomous driving deployment.

To address these challenges and enable comprehensive autonomous driving capabilities, the automotive industry has widely adopted high-definition (HD) maps, which represent a paradigmatic shift in cartographic technology characterized by three fundamental advantages over conventional navigation maps: enriched data content, enhanced spatial precision, and improved temporal freshness [13].

First, HD maps provide substantially richer data content compared to traditional road-level mapping. While conventional electronic maps record only basic road attributes such as geometry, grade, curvature, and directionality, HD maps encode comprehensive environmental details including elevated structures, guardrails, vegetation, road edge classifications, roadside landmarks, and precise lane marking typologies that enable fine-grained scene understanding [13]. Second, HD maps achieve centimeter-level spatial precision, typically maintaining absolute accuracy within one meter and relative accuracy of 10-20 centimeters, representing a significant improvement over standard definition (SD) maps with meter-level precision and commercial GPS systems with 5-meter accuracy. Third, HD maps maintain superior temporal freshness through quarterly update cycles, ensuring higher currency compared to

conventional navigation maps that undergo less frequent revision schedules.

This comprehensive environmental representation enables multiple critical autonomous driving functionalities. HD maps support precise map-based localization algorithms, ranging from classical feature matching and probabilistic filtering approaches to contemporary deep learning methodologies, achieving the sub-decimeter accuracy essential for safe autonomous navigation [6]. Beyond localization, HD maps facilitate enhanced perception through semantic priors that improve object detection in challenging conditions, enable advanced path planning with lane-level trajectory optimization, and support behavior prediction through detailed infrastructure modeling that anticipates likely trajectories of surrounding traffic participants [13].

However, the traditional mapping workflow for HD maps presents significant scalability constraints. This process typically involves specialized survey vehicles equipped with sophisticated sensor arrays (LiDAR, multi-view cameras, and high-precision inertial navigation systems), followed by extensive manual post-processing. The resultant workflow is both resource-intensive and time-consuming, creating difficulties in maintaining map currency across large geographical areas, particularly in urban environments subject to frequent infrastructural changes such as construction projects and road reconfiguration [1]. Given these limitations, leveraging existing HD maps as prior information to facilitate efficient map updates has emerged as a focal point in contemporary HD map research [44].

1.1 Evolution of Mapping and Localization

Simultaneous Localization and Mapping (SLAM) represents a foundational technology in mobile robotics and autonomous driving domains. The core objective of SLAM is to enable a robot or vehicle to concurrently construct an environmental map while estimating its own pose within that environment, with these dual processes reinforcing and optimizing each other [13]. Traditional SLAM approaches have predominantly relied on geometric features (such as corner points and line segments) [2] coupled with manually engineered observation models. Despite their effectiveness in structured environments, these classical approaches often struggle to disambiguate homogeneous structures such as repetitive poles or guardrails, limiting their reliability in complex scenarios.

Recent advances in deep learning have transformed both mapping and localization paradigms. Learning-based methods can extract rich semantic information which including lane markings, traffic signs, and curb boundaries, through end-to-end neural network architectures. These approaches frequently integrate extracted features into a BEV representation, which provides a top-down perspective of the surrounding environment. When combined with global priors contained in HD maps, these semantic features significantly enhance the robustness and interpretability of spatial matching operations [2, 40].

Unlike the unified framework of traditional SLAM, contemporary learning-based mapping and localization research has developed along largely parallel trajectories.

The first major research direction focuses on learning-based HD mapping technologies, which has evolved through distinct developmental phases. Early foundational approaches, ex-

emplified by HDMapNet [19], VectorMapNet [26], and MapTR [23], pioneered the paradigm of generating vectorized HD maps directly from raw sensor data without any prior knowledge integration. These seminal works established end-to-end learning pipelines that predict semantic map layers in BEV representations purely from current sensor observations. Subsequently, researchers began incorporating temporal knowledge through previous predictions and short-term observation buffers. Methods such as StreamMapNet [54] and PrevPredMap [31] demonstrated that leveraging recent mapping outputs and temporal perception continuity could significantly improve mapping stability and reduce frame-to-frame inconsistencies.

More recently, advanced approaches have integrated various forms of historical knowledge, including pre-existing map priors, neural map representations, and rasterized historical maps. Works like Neural Map Prior [50], HRMapNet [56], and P-MapNet [**jiang2024pmapnet**] have shown that incorporating long-term spatial knowledge can substantially enhance mapping accuracy and enable efficient map updating rather than complete reconstruction. However, across all evolutionary stages, these mapping-focused approaches typically rely on external GNSS/IMU systems for pose estimation, introducing significant localization errors that compromise mapping performance and limit practical deployment in GPS-challenging environments.

The second major research direction concentrates on learning-based HD map localization. This field has evolved from classical SLAM approaches [29], which are susceptible to cumulative drift over extended operational periods, toward methodologies that match real-time sensor observations with prior HD map knowledge. More recently, this evolution has led to deep learning-based pose regression systems that align BEV representations of current sensor inputs with stored semantic map embeddings [59]. Notable examples include SegLocNet [62], which unifies camera and LiDAR data through three-dimensional position encoding and cross-modal transformer architectures, and BEV-Locator [57], which employs transformer-based cross-modal association to align BEV features from multi-view images with semantic maps for direct vehicle pose regression. While these approaches achieve end-to-end learnable centimeter-level localization precision, they remain confined to the localization domain, lacking both unified evaluation metrics and systematic integration with mapping functionalities. Furthermore, their potential effectiveness in providing initial pose estimates for HD mapping remains largely unexplored.

The current separation between HD mapping and localization constitutes a critical research gap with significant implications for autonomous driving deployment [13]. Real-world autonomous systems require both continuously updated maps and robust localization within these evolving environmental representations. A fully integrated framework would ideally process potentially outdated prior maps alongside real-time sensor data streams to produce both updated HD maps and precise vehicle poses within an end-to-end architecture. Such integration would eliminate manual post-processing requirements, enable element-level interpretability (such as explicit detection of infrastructure changes), and provide synchronized map and pose data for downstream planning and control modules [1].

Nowadays, learning-based approaches to HD mapping and localization offer substantial advantages over traditional methodologies, including enhanced robustness, operational effi-

ciency, and scalability across dynamic and expansive environments. These benefits derive from the incorporation of semantic priors, end-to-end adaptive networks, multimodal sensor fusion capabilities, and vectorized representations which representing important evolutionary advancements beyond classical SLAM paradigms [40].

1.2 Problem Statement

Unlike traditional SLAM approaches that have inherently addressed localization and mapping as jointly optimized problems through hand-engineered feature extraction and geometric optimization techniques, currently learning-based HD mapping and localization research predominantly progress along parallel developmental trajectories: learning-based end-to-end HD mapping focuses on scalable and adaptive map creation methodologies, while HD map-based localization approaches emphasize robust pose estimation utilizing pre-constructed maps. Given the substantial resource requirements associated with comprehensive HD map production [32], efficient utilization of existing map information as prior knowledge for incremental map updates represents a primary industry concern.

Within this context, the traditionally distinct processes of mapping and localization become fundamentally interconnected, as research demonstrates that vehicle localization accuracy significantly impacts mapping performance. Despite this intrinsic relationship, there currently exists no unified framework capable of jointly learning to update HD maps while simultaneously performing localization within them, which constitutes a significant limitation that constrains system adaptability in dynamic environments. Additionally, the field lacks systematic exploration of the mutual interdependencies between localization accuracy and mapping performance, as well as comprehensive evaluation frameworks for assessing how these processes influence each other within HD map-based autonomous driving systems [13].

As a result we aim to bridging this gap between mapping and localization capabilities to enable scalable and adaptive autonomous systems capable of constructing and optimizing maps while determining precise poses in real-time operational contexts [13]. Compared to traditional SLAM methodologies, such integrated, learning-based approaches could substantially reduce dependence on costly manual mapping workflows while enabling deployment using conventional sensor configurations (cameras, LiDAR, GNSS, etc.). This integration would facilitate timely responses to environmental changes and dynamic conditions, establishing a foundation for truly end-to-end autonomous driving capabilities. Systematic investigation of existing methodologies coupled with the proposal of modular prototype architectures will illuminate critical research gaps and accelerate progress toward this objective.

1.3 Thesis Objectives

The primary aim of this thesis is to explore the integration of two recently developed but largely independent research domains: learning-based HD mapping and learning-based localization, to create systems capable of maintaining accuracy as operational environments evolve. To this end, the thesis pursues three interconnected objectives.

First, this research will provide a *comprehensive literature review* that situates classical HD map localization approaches, learning-based end-to-end HD mapping methodologies, and emerging learning-based localization techniques within a unified taxonomic framework. This review will systematically compare their respective methodologies, data requirements, evaluation metrics, and foundational assumptions. By elucidating their comparative strengths, limitations, and unexplored synergistic opportunities, this analysis will constitute a valuable reference resource for future research in learning-based HD mapping and localization domains.

Second, building upon these analytical insights, the thesis will develop a *modular architectural framework* for integrated mapping-localization pipelines. This design specification will define compatible representational formats, learning objectives, and system interfaces that enable mapping and localization processes to operate in either coupled or decoupled operational modes. This architectural exploration seeks to identify the optimal integration point where these traditionally separate tasks mutually reinforce rather than constrain each other.

Third, the research will conduct *practical validation* utilizing existing open-source implementations of learning-based HD mapping and localization systems in conjunction with widely adopted academic and industrial HD map datasets and evaluation metrics. This step will conclude analyses such as mapping quality assessment under varying initial pose error conditions. These experiments aim to identify potential integration points between mapping and localization technologies through systematic analysis, quantifying the impact of high-precision localization (sub-10 cm) on online map update capabilities. Additionally, these tests of existing methods will inform and validate the proposed modular architectural framework, with the scope of practical implementation determined by temporal and resource constraints. Collectively, these objectives establish a foundation for scalable and adaptive autonomous systems. By comprehensively examining the developmental trajectories of HD mapping and localization technologies, this thesis seeks to identify optimal integration strategies and provide architectural recommendations. The application of unified quantitative metrics and analysis will significantly advance autonomous driving capabilities that leverage prior HD map information, potentially contributing innovative approaches to next-generation autonomous transportation technologies.

Thesis Scope and Collaboration: This thesis, conducted as a collaborative initiative between Scania and the University of Twente under the supervision of Professor Maurice van Keulen (University of Twente) and industry experts Lena Wild and Rafael Valencia (Scania), aims to explore the convergence of these independently developing research domains. Through systematic analysis and design exploration, this work seeks to provide recommendations and architectural insights to inform the next generation of autonomous driving systems with particular emphasis on integrated mapping and localization capabilities.

Chapter 2

Background

This chapter provides a comprehensive review of high-definition (HD) maps in autonomous driving, including their origin, development challenges, and relevant literature on mapping and localization methodologies. The chapter also introduces fundamental concepts, datasets, sensor technologies, and evaluation metrics essential for understanding the research domain.

2.1 Sensor Technologies for Autonomous Driving

Achieving centimeter-level accuracy in autonomous driving systems relies fundamentally on effective *sensor fusion*, as no individual sensor can consistently deliver the required precision across all operational conditions. This section categorizes the relevant sensing modalities into two principal tiers: GNSS and IMU systems that provide an *absolute, high-rate motion baseline*, and exteroceptive sensors (LiDAR, RADAR, and cameras) that capture rich environmental structures for refinement against HD maps [58].

2.1.1 GNSS and IMU

Global Navigation Satellite Systems (GNSS), including GPS and Galileo, offer absolute positioning with approximately 5-10 meter accuracy under optimal conditions. While augmentation technologies like Differential GPS (DGPS), Real-Time Kinematic(RTK) and Dual-frequency can improve precision to sub-meter or even centimeter-level accuracy, these signals remain vulnerable to multipath effects and signal blockages in urban environments and enclosed spaces such as tunnels. Inertial Measurement Units (IMU) complement GNSS by integrating acceleration and angular rate measurements at kilohertz frequencies, providing smooth short-term motion estimates. However, IMU measurements accumulate errors over time, necessitating periodic corrections from absolute sensors or alignment with static map features using LiDAR or camera data. In contemporary autonomous driving systems, GNSS primarily serves as an initial pose estimator, after which more precise sensors or map-matching algorithms assume the localization task [57, 62, 28].

2.1.2 LiDAR

Light Detection and Ranging(Lidar) systems emit laser pulses and measure their return times to construct three-dimensional point clouds with centimeter-level accuracy. Mechanical LiDAR units typically provide full 360°coverage, creating ideal conditions for fine-grained registration against HD maps. Lane boundaries, curbs, poles, and road edges appear as distinct features that facilitate precise scan-to-map correspondence. Despite their advantages, LiDAR systems present certain limitations, including high data rates, sensitivity to adverse weather conditions (heavy rain or fog), and historically higher costs. However, recent advancements in solid-state LiDAR technology are progressively addressing these constraints [13].

2.1.3 RADAR

Automotive Radio Detection and Ranging(RADAR) systems measure both range and radial velocity through Doppler shift analysis. RADAR technology demonstrates reliable performance across challenging environmental conditions, including rain, snow, dust, and glare, while offering a wide field of view at relatively low cost [32]. The primary limitations of RADAR include comparatively coarse angular resolution and sensitivity to target reflectivity characteristics, which necessitate careful filtering when utilizing RADAR data for ego-motion estimation or map alignment.

2.1.4 Camera Systems

Camera configurations, whether monocular, stereo, or surround-view, capture high-resolution imagery suitable for visual odometry and semantic landmark detection, such as traffic signs and lane markings. Cameras represent cost-effective and lightweight sensor options but exhibit performance degradation in low-light conditions and adverse weather, while also imposing significant computational requirements [8]. Recent advancements in object detection and semantic segmentation now enable camera-based systems, when integrated with HD maps, to achieve localization accuracies previously exclusive to LiDAR-based approaches, provided sufficient environmental texture and illumination conditions are present.

In contemporary autonomous driving architectures, localization typically begins with a GNSS/IMU prior estimate, then converges to centimeter-level accuracy by aligning LiDAR, RADAR, or camera observations with HD map representations. Each sensing modality compensates for the limitations of others: GNSS provides global coverage but fails in signal-obstructed environments; IMU bridges short outages but accumulates drift; LiDAR offers precise geometry but at higher cost; RADAR maintains functionality in adverse weather but with limited resolution; and cameras contribute rich semantic information at minimal cost. Effective integration of these complementary sensing modalities represents a critical factor in developing robust, cost-effective localization solutions for autonomous vehicles.

2.2 High-Definition Maps

The concept of High-Definition (HD) maps emerged in 2010 at a Mercedes-Benz research facility in Stuttgart. Engineers envisioned cartographic representations with unprecedented detail and accuracy that could function as "virtual sensors," providing autonomous vehicles with comprehensive three-dimensional environmental information extending beyond the range of onboard sensors. The 2013 Bertha Drive project, which featured an autonomous Mercedes-Benz vehicle navigating Germany's historic Bertha Benz route using HD maps developed by HERE, demonstrated the viability of this approach. Throughout the 104-kilometer journey, the vehicle successfully navigated complex interchanges, roundabouts, and urban environments by matching real-time sensor data with map-encoded lane geometry and semantic landmarks, validating the role of HD maps as an essential component for safe autonomous operation [13].

HD maps differ fundamentally from conventional mapping systems in several critical aspects. First, they capture road geometry-including lane centerlines, boundaries, and intersection configurations-with 5-10 centimeter accuracy. Second, they incorporate comprehensive semantic attributes: each lane contains annotations regarding permissible driving routes, speed profiles, and turning zones, while traffic signs, curb heights, pedestrian crossings, and roadside infrastructure are encoded with precise three-dimensional coordinates. Finally, these maps organize information within a graph structure that defines lane-level connectivity and regulatory constraints, enabling route planning and behavior prediction algorithms to operate with unprecedented granularity [44].

These distinctive features enable transformative capabilities for autonomous driving systems [13]. High-precision localization fuses real-time sensor data with the map's three-dimensional point cloud and semantic landmarks to achieve lateral positioning accuracy within centimeters-essential for executing safe lane-level maneuvers. Enhanced perception leverages map-based prior knowledge to improve object detection and classification in challenging scenarios such as faded road markings or reduced visibility conditions. Advanced route and trajectory planning utilizes the lane connectivity graph to generate kinematically optimal paths that conform to speed limits, curvature constraints, and turning regulations. Finally, behavior prediction benefits from detailed lane and semantic models, allowing the system to anticipate likely trajectories of surrounding road users based on available infrastructure elements such as bicycle lanes and pedestrian crossings.

Despite significant research and development efforts from both academic and industrial sectors to advance HD mapping technologies, several unresolved challenges continue to impede their full potential for autonomous mobility applications [13]. These challenges can be categorized as follows:

- **Data Collection**: The acquisition of HD map data represents a time-intensive and resource-demanding process. Standard collection procedures involve deploying specialized vehicles equipped with GNSS, IMU, LiDAR, and camera arrays to capture detailed environmental information [40]. The integration of multiple sensor inputs, combined with variable environmental conditions, significantly influences data quality.

Furthermore, the economic aspect of HD map generation presents substantial barriers to widespread implementation, as large-scale mapping operations require extensive fleets of specialized vehicles equipped with high-precision sensing systems. Although consumer-grade sensors might theoretically support HD mapping applications, their deployment necessitates more sophisticated algorithmic approaches to compensate for reduced sensor fidelity[13].

- **Data Communication**: Efficient transmission of mapping data from collection platforms to processing centers, and subsequently to autonomous vehicles, presents significant technical challenges. Mapping vehicles generate voluminous multi-sensor datasets that require processing for map construction and maintenance. Real-time processing of data streams from multiple mapping vehicles imposes considerable demands on computational resources and transmission bandwidth, creating bottlenecks in system development [32].

- **Data Processing**: The extraction of essential elements and features for HD map construction involves complex computational processes, particularly for extensive geographical areas [3]. This task requires aggregating and aligning data from diverse sources while ensuring spatial accuracy and temporal currency [49]. When multiple mapping vehicles participate in data collection, precise temporal synchronization becomes critical to prevent misalignment issues. Currently, GPS-generated pulse-per-second (PPS) signals represent the standard approach for synchronizing onboard sensor systems [32].

- **Map Maintenance**: The continuous updating of HD maps in response to environmental changes-including construction activities, road reconfigurations, and infrastructure modifications-requires frequent data collection and processing. Failure to incorporate timely updates regarding lane closures or new traffic signage could potentially lead to inappropriate vehicle behavior. Maintaining current, accurate map information is essential for safe autonomous operation and represents a more efficient approach than generating entirely new maps. However, research on efficient map maintenance remains limited, with few studies and restricted dataset availability [18, 39]. Recent developments have begun to explore end-to-end explainable HD map update methodologies [45].

- **Data Privacy and Security**: HD maps frequently contain sensitive information regarding critical infrastructure, raising important privacy and security considerations. Ensuring appropriate data protection against unauthorized access or misuse represents a significant challenge. Regulatory frameworks vary across jurisdictions, and mapping activities often intersect with national security concerns, restricting cross-border data sharing [13]. Conducting research and operational implementations within applicable legal parameters remains essential for responsible development.

Addressing these multifaceted challenges requires interdisciplinary collaboration across computer science, geographic information systems, traffic engineering, and data security domains to advance HD mapping technologies. The research presented in this thesis contributes to addressing these challenges through the development of integrated end-to-end approaches

for real-time localization and mapping, potentially enhancing the reliability and accessibility of autonomous driving technologies.

2.3 Datasets

Research datasets play a pivotal role in the development and evaluation of mapping and localization methodologies for autonomous vehicles. The selection of appropriate datasets significantly influences the effectiveness and generalizability of proposed models. This section introduces prominent datasets frequently employed in learning-based mapping and localization research.

- **nuScenes** [4]: The nuScenes dataset encompasses approximately 1.4 million camera images, 390,000 LiDAR scans, and 1.4 million millimeter-wave radar scans. Manual annotations provide 3D bounding boxes for over 1.4 million objects, captured at a 2Hz frequency across 40,000 keyframes, covering diverse urban and suburban environments in Boston and Singapore. Additionally, the dataset includes object-level attribute annotations such as visibility conditions, activity states, and pose information.

  The dataset supports a comprehensive range of autonomous driving tasks, including 3D object detection, multi-object tracking, motion prediction, LiDAR segmentation, and panoptic segmentation and tracking. Its applications extend to multi-agent prediction, pedestrian localization, weather augmentation, and mobile point cloud prediction research. The dataset's focus on challenging urban environments across geographically distinct locations enhances its value for developing robust autonomous driving systems.

- **Argoverse** [7, 47]: The Argoverse dataset contains more than 300,000 frames of sensor data and over 200,000 3D object annotations. The dataset has evolved from Argoverse 1 to Argoverse 2, expanding both its scale and feature richness. While Argoverse 1 data collection occurred in Pittsburgh and Miami, Argoverse 2 expanded coverage to six U.S. cities: Austin, Detroit, Miami, Pittsburgh, Palo Alto, and Washington, D.C.

  A notable component of the Argoverse ecosystem is the Map Change dataset [18], which constitutes part of the Argoverse 2 maps and represents the first public dataset specifically designed for high-precision map change detection. This resource addresses the critical challenge of identifying inconsistencies between sensor data and map information resulting from real-world environmental changes. Researchers utilize simulated data (synthetically modified maps) to train models for change detection in both BEV and ego-vehicle perspectives. The dataset focuses on permanent map modifications such as lane geometry and pedestrian crossings, while excluding temporary changes such as construction activities. This resource enables the development of algorithms capable of detecting outdated HD map information due to environmental evolution.

  However this dataset lacks complete map priors for training and ground-truth maps for testing, forcing reliance on synthetic priors generated through random noise and simple rules that create a significant gap between artificial perturbations and structured real-world road changes, preventing effective model generalization to actual scenarios.

ArgoTweak [46] is the first dataset to provide complete realistic map priors with systematic change annotations. This dataset is derived from the Argoverse 2 Map Change Dataset. ArgoTweak introduces a bijective change mapping framework that decomposes complex road modifications into traceable atomic changes (geometry, markings, insertions, etc.), provides the first complete prior-sensor-ground truth triplet dataset, and implements fine-grained evaluation metrics that distinguish between map preservation and updating capabilities, significantly reducing the simulation-to-reality performance gap.

- **KITTI** [14]: The KITTI dataset contains extensive traffic scene recordings, including more than 200,000 stereo images with corresponding point cloud data. Data collection encompassed urban, suburban, and rural environments in Karlsruhe and surrounding communities in Germany. The dataset provides ground truth pose information recorded across urban environments and highway segments. Unlike previously mentioned datasets, KITTI utilizes OpenStreetMap (OSM) as its mapping format, commonly employed as a standard-definition map in positioning research [35].

  KITTI has established itself as a benchmark standard for evaluating diverse computer vision tasks related to autonomous driving. Its evaluation domains include stereo vision (stereo matching, disparity estimation), optical flow estimation, visual odometry/SLAM, 3D object detection, 3D object tracking, scene understanding, road and lane detection, semantic segmentation, and depth completion/prediction.

2.4 Evaluation Metrics

Appropriate metrics are essential for quantifying the performance of mapping and localization algorithms. Different evaluation criteria assess accuracy, robustness, and efficiency characteristics. This section introduces widely adopted metrics in autonomous vehicle localization research:

Relative Pose Error (RPE) RPE quantifies the *local drift* of a trajectory by comparing estimated incremental motion between two time points against ground-truth motion over the same interval [38]. Let $\mathbf{T}_i \in \mathrm{SE}(3)$ represent the ground-truth pose at time i, and $\hat{\mathbf{T}}_i$ its estimate. For a fixed time lag $\Delta > 0$:

$$\mathbf{E}_{i,\Delta}^{\mathrm{rpe}} = \left(\mathbf{T}_i^{-1}\mathbf{T}_{i+\Delta}\right)^{-1}\left(\hat{\mathbf{T}}_i^{-1}\hat{\mathbf{T}}_{i+\Delta}\right). \tag{2.1}$$

The translational RPE is calculated as $\|\operatorname{trans}(\mathbf{E}_{i,\Delta}^{\mathrm{rpe}})\|_2$ (meters), while the rotational RPE is $\angle(\operatorname{rot}(\mathbf{E}_{i,\Delta}^{\mathrm{rpe}}))$ (degrees or radians). Statistical measures including the mean, RMSE, or specified percentiles across all valid indices i provide aggregate performance indicators.

Absolute Pose Error (APE) APE measures the *global consistency* of an estimated trajectory relative to ground truth. After aligning the estimate with a similarity transform $\mathbf{S} \in \mathrm{Sim}(3)$ (typically via the Umeyama algorithm):

$$\mathbf{E}_i^{\text{ape}} \;=\; (\mathbf{T}_i)^{-1}\,\mathbf{S}\,\hat{\mathbf{T}}_i. \tag{2.2}$$

The translational APE is computed as $\|\operatorname{trans}(\mathbf{E}_i^{\text{ape}})\|_2$ (meters), while the rotational APE is $\angle(\operatorname{rot}(\mathbf{E}_i^{\text{ape}}))$. Since APE accumulates drift over time, it is particularly relevant for evaluating long-range navigation performance.

Longitudinal/Latitudinal/Orientation Recall (LLO-Recall). For safety-critical applications, determining *how frequently* errors remain *within* specified tolerances is often necessary. The per-axis indicator function is defined as:

$$\mathbb{I}_i^x(\tau_x) \;=\; \begin{cases} 1, & \text{if } |\hat{x}_i - x_i| < \tau_x, \\ 0, & \text{otherwise,} \end{cases}$$

with analogous definitions for $\mathbb{I}_i^y(\tau_y)$ (latitude) and $\mathbb{I}_i^\theta(\tau_\theta)$ (heading). The corresponding recall metrics are calculated as:

$$\text{Recall}_x = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_i^x(\tau_x), \quad \text{Recall}_y = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_i^y(\tau_y), \quad \text{Recall}_\theta = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_i^\theta(\tau_\theta), \tag{2.3}$$

expressed within the range $[0, 1]$. Typical threshold values include $\tau_x, \tau_y \in \{0.1, 0.2\}$ m and $\tau_\theta \in \{1^\circ, 2^\circ\}$.

Mean Average Precision (mAP) For evaluating the quality of HD map generation, particularly vectorized map elements such as lanes, boundaries, and intersections, mean Average Precision (mAP) serves as the primary metric [43]. This metric evaluates the spatial accuracy of detected map elements by measuring the overlap between predicted and ground-truth vectorized representations.

For each map element class c (e.g., lane dividers, road boundaries, pedestrian crossings), the Average Precision $\text{AP}_c(\tau)$ at distance threshold τ is calculated as:

$$\text{AP}_c(\tau) \;=\; \int_0^1 \text{Precision}_c(r, \tau)\, dr \tag{2.4}$$

where $\text{Precision}_c(r, \tau)$ represents the precision at recall level r for class c when using distance threshold τ. A predicted map element is considered a true positive if its Chamfer distance to the nearest ground-truth element of the same class is below the threshold τ.

The mean Average Precision across all map element classes is then computed as:

$$\text{mAP}(\tau) \;=\; \frac{1}{C}\sum_{c=1}^{C}\text{AP}_c(\tau) \tag{2.5}$$

where C is the total number of map element classes. In HD mapping evaluation, mAP is commonly reported at multiple distance thresholds to assess mapping precision at different scales:

$$\text{mAP} \;=\; \frac{1}{3}\big[\text{mAP}(0.5) + \text{mAP}(1.0) + \text{mAP}(1.5)\big] \tag{2.6}$$

The standard evaluation thresholds are $\tau \in \{0.5\,\text{m}, 1.0\,\text{m}, 1.5\,\text{m}\}$, reflecting the precision requirements for different autonomous driving applications. The 0.5m threshold evaluates fine-grained mapping accuracy essential for precise lane-keeping, while the 1.5m threshold assesses broader road structure detection suitable for path planning applications.

Mean Absolute Error (MAE) For a scalar error sequence $\{e_i\}_{i=1}^{N}$:

$$\text{MAE} \;=\; \frac{1}{N}\sum_{i=1}^{N}|e_i|. \tag{2.7}$$

MAE exhibits linear scaling with errors, providing enhanced interpretability when large outliers are possible but not dominant within the distribution.

Root Mean Squared Error (RMSE)

$$\text{RMSE} \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}e_i^2}. \tag{2.8}$$

RMSE imposes quadratic penalties on large errors, demonstrating greater sensitivity to outliers compared to MAE. Both MAE and RMSE metrics may be reported for translational (meters) or rotational (degrees/radians) errors by substituting the corresponding error terms.

Unless otherwise specified, all statistical metrics presented in this research utilize the complete trajectory length (N time steps) and adhere to SI unit conventions.

Chapter 3

Key Techniques in Each Domain

This chapter provides a comprehensive analysis of the state-of-the-art methodologies in both HD map-based localization and HD mapping domains. The analysis is structured to systematically examine the evolution, architectural innovations, and performance characteristics of each field, establishing the foundation for understanding potential integration opportunities. The localization domain encompasses both HD map-based approaches that achieve centimeter-level accuracy through sophisticated semantic matching, and SD map-based methods that provide meter-level positioning while maintaining broader operational robustness. The mapping domain focuses on online HD map construction techniques categorized by their utilization of prior information, ranging from single-frame approaches to sophisticated temporal and historical integration strategies. Through detailed experimental analysis using standardized datasets and metrics, this chapter quantifies the trade-offs between accuracy, computational efficiency, and robustness across different architectural paradigms. The comparative evaluation reveals fundamental insights into the complementary strengths and limitations of each approach, informing the subsequent architectural framework design for integrated mapping-localization systems.

3.1 Map-Based Localization Review

3.1.1 HD map-based Localization Methods

Visual localization represents a fundamental challenge in autonomous driving systems, where precise vehicle positioning is crucial for safe navigation and path planning. Recent advances in deep learning have enabled sophisticated end-to-end approaches that leverage Bird's Eye View (BEV) representations and semantic map information to achieve centimeter-level accuracy [57, 15, 28].

The evolution from traditional perception-matching-optimization pipelines to end-to-end learning paradigms has demonstrated promising improvement in both computational efficiency and localization accuracy. This thesis provides a comprehensive evaluation of three state-of-the-art methods that exemplify different architectural strategies for visual localization.

Traditional methods like Iterative Closest Point (ICP) which are applied in Zhao et al. ICIP 2024 [60] represents a classical geometric registration approach that iteratively minimizes the distance between corresponding points in LiDAR point clouds and HD map representations. The algorithm alternates between establishing point correspondences and computing optimal rigid-body transformations until convergence. While computationally efficient and conceptually straightforward, ICP suffers from sensitivity to initialization and lacks explicit uncertainty quantification, limiting its robustness in challenging urban environments with sparse or ambiguous geometric features.

BEV-Locator [57] establishes visual semantic localization as a comprehensive end-to-end learning paradigm, requiring only simple pose offset supervision generated from vehicle trajectories with raw images and semantic maps [57]. This streamlined supervision strategy enables efficient training dataset generation while achieving sophisticated cross-modal understanding capabilities between BEV representations and semantic map information.

The transformer-based architecture enables efficient centimeter-level positioning through sophisticated BEV-to-semantic map multimodal fusion perception capabilities. The end-to-end framework eliminates traditional perception-matching-optimization pipeline complexities while maintaining differentiable optimization throughout the entire system.

The Decoupled BEV Neural Matching approach introduces a fundamental architectural innovation by decomposing pose estimation into separate yaw, longitudinal, and lateral components [28]. This decoupling strategy enables independent sampling and matching for each dimension, substantially improving both computational efficiency and system interpretability compared to traditional joint optimization approaches.

The decoupled matcher effectively reduces search space complexity from cubic to linear scaling, achieving a remarkable 68.8% reduction in inference memory usage while maintaining competitive accuracy. The architectural design leverages advanced BEV perception networks to enhance feature quality despite relying solely on camera inputs.

EgoVM [15] demonstrates exceptional performance through sophisticated transformer decoder architectures combined with lightweight vectorized map representations [15]. This design preserves essential geometric relationships while reducing computational overhead, facilitating more precise pose estimation capabilities compared to alternative approaches.

The method's architectural sophistication enables robust geometric constraint learning through advanced cross-modal matching techniques, with the incorporation of LiDAR point clouds further enhancing localization accuracy in multi-modal configurations.

U-ViLAR [21] introduces an uncertainty-aware visual localization framework that addresses the fundamental challenges of perception and localization uncertainty in autonomous driving systems. The method employs a two-stage architecture combining Perceptual Uncertainty-guided Association with Localization Uncertainty-guided Registration to achieve robust and accurate positioning across different map formats.

The Perceptual Uncertainty-guided Association component models perceptual degradation effects such as occlusions and missing semantic elements by incorporating uncertainty estimation into cross-modal association between visual BEV features and map representations. This approach utilizes both global association through uncertainty-aware similarity

matrices and local association via contrastive learning within sampled windows to establish fine-grained correspondences.

The Localization Uncertainty-guided Registration stage constructs a 3D solution space centered around coarse pose estimates, leveraging joint probability distributions across all degrees of freedom to enable error correction through probabilistic-guided pose registration. This design addresses multimodal distribution challenges and potential degeneracy in specific degrees of freedom that traditional coarse-to-fine approaches often overlook.

Experimental Results

Table 3.1 presents the comparative performance analysis of the evaluated methods on the nuScenes dataset. The results demonstrate significant variations in localization accuracy across different dimensional metrics and sensor configurations. Note that ICP methods use APE (Absolute Pose Error)(2.2) metric which combines longitudinal and lateral errors, while other methods report MAE (Mean Absolute Error)(2.7) for individual dimensions.

TABLE 3.1: HD Map-Based Localization Performance on nuScenes Dataset

Method	**Long./APE (m)**	**Lat. (m)**	**Yaw (°)**	**Init. Pose Error**	**Metircs**
ICP	0.39	–	–	–	APE
Zhao et al. [60]	0.34	–	–	Defined, not quantified	APE
EgoVM [15]	0.151	0.047	0.092	±3 m, ±3°, both axes	MAE
BEV-Locator [57]	0.178	0.076	0.510	±2 m (long), ±1 m (lat), ±2°	MAE
U-ViLAR [21]	0.140	0.040	0.075	±2 m (long), ±2 m (lat), ±2°	MAE
Decoupled BEV [28]	0.19	0.13	0.39	±2 m (long), ±2 m (lat), ±2°	MAE

The experimental results reveal several key insights about the comparative effectiveness of different architectural approaches:

State of the Art Performances: EgoVM demonstrates exceptional performance across all dimensional metrics, achieving 0.151 meters longitudinal, 0.047 meters lateral, and 0.092°yaw errors that consistently surpass other methods. The exceptional lateral precision of 0.047 meters represents the most accurate performance among evaluated approaches. U-ViLAR [21] achieves superior longitudinal precision (0.140 meters) and excellent yaw estimation (0.075°) while maintaining competitive lateral performance (0.040 meters) under moderate perturbation conditions (±2m, ±2°). The method's strength lies in explicit uncertainty modeling that addresses both perceptual degradation and localization ambiguity, providing uncertainty quantification essential for safety-critical applications alongside high localization accuracy. The inputs

BEV-Locator Baseline Performance: BEV-Locator achieves excellent pose accuracy with lateral and longitudinal position errors remaining competitive. However, the heading direction prediction exhibits significant variability, manifesting in an overall yaw error of 0.510°, which is substantially higher than other methods.

Decoupled BEV Efficiency Trade-offs: The Decoupled BEV approach demonstrates effective balance between computational efficiency and localization precision. The decoupled version achieves decimeter-level precision with Mean Absolute Error values of 0.19m longitudinal, 0.13m lateral, and 0.39°, representing a reasonable compromise for practical deployment scenarios.

Traditional Methods Comparison: The traditional ICP and Zhao et al. ICIP 2024 [60] methods, using APE (Absolute Pose Error) metric, achieve 0.39m and 0.34m respectively. While these methods do not provide separate dimensional analysis, their overall positioning accuracy demonstrates competitive performance, with [60] achieving better accuracy than ICP. While Zhao et al. ICIP 2024 [60] also performs BEV perception similar to the these learning-based methods (BEV-Locator [57], Decoupled BEV [28], and EgoVM [15]) U-Vilar [21], a fundamental distinction lies in their pose estimation strategies. Zhao et al. ICIP 2024 [60] employs a traditional two-stage pipeline where BEV feature extraction and pose optimization are performed sequentially with separate, non-differentiable optimization procedures. In contrast, the three learning-based methods implement fully end-to-end(E2E) architectures where pose estimation is integrated into the neural network through differentiable operations. This E2E design enables joint optimization of feature extraction and pose estimation during training, allowing the network parameters to be automatically tuned for optimal localization performance. Consequently, the learning-based approaches achieve superior robustness and accuracy by learning task-specific feature representations that are directly optimized for the localization objective, rather than relying on hand-crafted feature matching and separate geometric optimization steps.

All these E2E approaches converge on transformer or decoder architectures for implementing cross-modal matching between BEV image features and map representations. This architectural convergence reflects community recognition that learned attention mechanisms provide superior capabilities for establishing complex cross-modal correspondences compared to handcrafted feature matching approaches.

A critical limitation emerges in the conservative perturbation assumptions of 2-3 meters employed across these methods, which inadequately simulate realistic GNSS deviations that frequently exceed 10 meters in challenging urban environments. This discrepancy between experimental validation conditions and actual deployment uncertainties suggests that while these methods demonstrate excellent refinement capabilities for relatively accurate initial estimates, their robustness for global relocalization scenarios remains inadequately characterized.

3.1.2 SD Map-Based Localization Results Analysis

Standard Definition (SD) map-based visual localization emerges as a critical paradigm for large-scale autonomous driving deployment, where the prohibitive costs and limited coverage of High Definition (HD) maps necessitate alternative approaches. Unlike HD maps that require expensive LiDAR surveys and manual annotation, SD maps leverage freely available resources such as OpenStreetMap (OSM) while achieving competitive localization accuracy through sophisticated neural architectures [35, 48, 62].

The evolution from traditional feature-matching pipelines to E2E learning frameworks has demonstrated remarkable progress in bridging the semantic gap between visual observations and lightweight map representations. This comprehensive evaluation examines fmy state-of-the-art methods that represent distinct architectural paradigms for SD map-based localization, showcasing the transition from single-modal to multi-modal approaches and from

regression-based to matching-based pose estimation strategies.

SegLocNet [62] establishes a groundbreaking multi-modal localization framework that combines surround-view images with LiDAR point clouds to construct comprehensive Bird's Eye View (BEV) semantic representations. The method's core innovation lies in its exhaustive matching strategy that directly correlates predicted BEV segmentation masks with prior map representations, eliminating the limitations of regression-based pose estimation approaches.

The unified map representation design enables seamless adaptation between different map configurations without architectural modifications, while the multi-modal sensor fusion significantly enhances robustness in challenging urban environments. The exhaustive matching paradigm provides superior generalization capabilities compared to regression-based alternatives, particularly evident in cross-dataset evaluation scenarios.

MapLocNet [48] introduces a sophisticated coarse-to-fine feature registration framework that leverages transformer architectures for hierarchical BEV-to-map alignment. The method's architectural innovation centers on decomposing pose estimation into sequential registration stages, where coarse alignment provides initial pose estimates that guide subsequent fine-grained refinement processes.

The neural localization modules employ learned attention mechanisms to establish complex correspondences between visual BEV features and neural map representations. While achieving competitive accuracy metrics, the regression-based pose estimation strategy exhibits sensitivity to training distribution shifts, limiting generalization performance in unseen environments compared to matching-based alternatives.

U-BEV [5] demonstrates significant architectural advancement through height-aware BEV representation learning that explicitly models 3D scene geometry while maintaining computational efficiency. The method's key contribution lies in discretizing height information rather than depth, substantially reducing computational complexity while preserving essential geometric constraints for accurate localization.

The height-aware projection mechanism enables superior scene understanding in urban environments with significant vertical structure variations. However, the method's limitation to 2-DoF pose estimation (position only) restricts its applicability in scenarios requiring full 3-DoF pose recovery including orientation estimation.

OrienterNet [35] pioneers neural matching between monocular images and OSM representations through learned Bird's Eye View inference and probabilistic pose estimation. The method's significance lies in demonstrating that lightweight planimetric maps contain sufficient geometric constraints for accurate visual localization when combined with sophisticated neural architectures.

The probabilistic pose estimation framework enables multi-modal likelihood distributions that capture localization uncertainty, particularly valuable for sensor fusion applications. The method's reliance on single-camera input provides computational advantages but limits performance in visually ambiguous environments compared to multi-modal approaches.

Experimental Results

Table 3.2 presents comprehensive performance analysis of the evaluated methods on the nuScenes dataset, revealing significant variations in localization accuracy across different sensor configurations and map representations. The results demonstrate the progressive improvements achieved through architectural innovations and multi-modal sensor integration.

TABLE 3.2: SD Map-Based Localization Performance on nuScenes Dataset

Position Metrics					
Method	**Sensors**	**Pos@1m (%)**	**Pos@2m (%)**	**Pos@5m (%)**	**APE (m)**
OrienterNet [35]	Single-Cam	21.73	35.36	50.53	14.79
U-BEV [5]	Multi-Cam	16.89	41.60	71.33	–
MapLocNet [48]	Multi-Cam	20.10	45.54	77.70	–
SegLocNet-SD [62]	Multi-Cam	35.63	57.98	74.55	8.15
SegLocNet-NM [62]	Multi-Cam	48.23	59.22	61.55	13.60
SegLocNet-HD [62]	Multi-Cam	59.08	76.04	84.25	5.30
Orientation Metrics					
Method	**Sensors**	**Ori@1°(%)**	**Ori@2°(%)**	**AOE (°)**	**Init. Pose Error**
OrienterNet [35]	Single-Cam	32.78	47.51	46.04	±20 m (x, y), ±10°
U-BEV [5]	Multi-Cam	–	–	–	[0, 100] m, none
MapLocNet [48]	Multi-Cam	58.61	84.10	–	±30 m (x, y), ±30°
SegLocNet-SD [62]	Multi-Cam	38.58	64.98	19.68	±32 m (x, y), none
SegLocNet-NM [62]	Multi-Cam	31.09	57.25	40.79	±32 m (x, y), none
SegLocNet-HD [62]	Multi-Cam	63.19	84.55	10.11	±32 m (x, y), none

- **SegLocNet-SD:** Uses SD map from OpenStreetMap for localization
- **SegLocNet-HD:** Uses HD map combined with SD maps for enhanced localization precision
- **SegLocNet-NM:** Employs Neural Maps approach processes rasterized SD maps through U-Net architecture to generate neural map representations
- **U-BEV:** Performs 2-DOF pose estimation (position only), hence orientation metrics are not applicable
- "-" indicates metrics not reported in the original studies

The experimental results reveal several critical insights about the comparative effectiveness of different architectural paradigms and sensor configurations:

Multi-modal Sensor Fusion Superiority: SegLocNet-SD demonstrates exceptional performance advantages through multi-modal sensor integration, achieving 35.63% recall at 1-meter accuracy compared to 21.73% for single-camera OrienterNet. This represents a substantial improvement in precision localization capabilities (from 21.73% to 35.63%, an absolute gain of 13.90 percentage points), highlighting the fundamental importance of complementary sensor information for robust pose estimation in challenging urban environments.

Hybrid Map Representation Impact: The revolutionary SD+HD fusion strategy in SegLocNet-HD demonstrates optimal resource utilization by selectively incorporating high-precision HD road geometry (drivable areas) while leveraging comprehensive SD map coverage for buildings and Points of Interest (POI). This strategic combination achieves 59.08% recall at 1-meter accuracy compared to 35.63% for pure SD maps, representing a 66% performance improvement. Critically, the hybrid approach maintains cost-effectiveness by utilizing expensive HD mapping resources only for essential road geometry while relying on freely available OSM data for contextual information, presenting a scalable solution for large-scale autonomous driving deployment.

Architectural Strategy Trade-offs: MapLocNet achieves superior 5-meter recall performance (77.70%) and competitive orientation accuracy (58.61% at 1°) through its coarse-to-

fine registration strategy. However, the method's regression-based approach exhibits reduced performance at fine-grained precision levels (20.10% at 1-meter), illustrating the trade-off between different pose estimation paradigms.

Computational Efficiency Considerations: U-BEV demonstrates balanced performance with 71.33% recall at 5-meter accuracy while maintaining computational efficiency through height-aware BEV representations. The method's limitation to position-only estimation restricts comprehensive pose recovery but enables deployment in resource-constrained scenarios.

Generalization Capability Assessment: SegLocNet's exhaustive matching strategy exhibits superior cross-dataset generalization compared to regression-based alternatives, with consistent performance improvements across both nuScenes and Argoverse datasets. This robustness stems from the method's ability to capture multi-modal pose distributions rather than single-point estimates.

All evaluated methods converge on transformer-based architectures for implementing cross-modal correspondence learning between BEV representations and map information. This architectural consensus reflects the community's recognition that learned attention mechanisms provide superior capabilities for establishing complex semantic-geometric relationships compared to traditional handcrafted feature matching approaches.

The fundamental distinction between HD and SD map-based localization lies in their precision-robustness characteristics and operational assumptions. HD map-based methods typically achieve **centimeter-level accuracy** (1-5 cm Mean Absolute Error) but operate under stringent constraints of small initial pose uncertainties (2-5 meters), requiring high-quality prior localization estimates. Conversely, SD map-based approaches sacrifice precision to **meter-level accuracy** (0.1-1 meter MAE) while demonstrating superior robustness to large initial pose deviations (20-100 meters), better reflecting real-world GNSS degradation scenarios in urban environments.

Remarkably, despite these substantial performance and operational differences, both paradigms converge on similar **architectural foundations**-transformer-based BEV representations and cross-modal attention mechanisms-suggesting that the core technical challenge lies in semantic-geometric correspondence learning rather than fundamental architectural innovations. SegLocNet's hybrid SD+HD approach strategically positions itself in the **precision-robustness spectrum**, achieving decimeter-level accuracy (5-10 cm MAE) while maintaining moderate robustness to initial pose uncertainties, demonstrating the viability of selective map quality integration.

The larger-scale prior map coverage employed in SD map-based methods (typically 200-500 meter map tiles vs. 50-100 meter tiles in HD approaches) provides additional contextual constraints that enhance localization disambiguation in visually ambiguous scenarios, offsetting some precision limitations through improved semantic understanding of the broader spatial context.

Strategic Innovation in Hybrid Map Fusion: The emergence of SegLocNet's SD+HD fusion strategy represents a paradigmatic shift toward **selective resource utilization** in autonomous driving localization. This approach demonstrates that strategic

integration of high-precision HD road geometry with comprehensive SD contextual information achieves superior performance while maintaining economic viability. The hybrid strategy addresses the fundamental scalability limitations of pure HD map approaches while overcoming the precision constraints of pure SD map solutions, establishing a promising framework for future large-scale deployment scenarios.

Furthermore, the computational complexity scaling of multi-modal approaches presents challenges for real-time deployment, with SegLocNet requiring sophisticated sensor synchronization and processing pipelines that may limit practical adoption in cost-sensitive applications. The trade-off between localization accuracy and system complexity remains a critical consideration for autonomous driving platform integration.

3.2 HD Mapping Review

Online HD mapping has emerged as a critical component for autonomous driving systems, enabling real-time generation of high-definition maps from onboard camera sensors. The integration of prior information has proven to be a pivotal factor in enhancing mapping performance across different temporal and spatial scales. This section provides a comprehensive analysis of visual-based HD mapping methods categorized by their utilization of prior information: methods without prior knowledge, those leveraging temporal priors, and approaches incorporating long-term historical map priors.

3.2.1 No Prior Information Methods

Methods in this category rely solely on current camera observations without incorporating any prior knowledge or historical information. These approaches form the foundation of online HD mapping and demonstrate the baseline capabilities of various algorithmic frameworks for vectorized map construction.

CNN-Based Mapping Methods

CNN-based methods utilize convolutional neural networks to process image or BEV features hierarchically for vectorized map construction. HDMapNet [19], as the pioneering work, employs a fully convolutional network with three branches for semantic segmentation, instance embedding, and direction prediction. Despite its foundational importance, HDMapNet achieves only 23.0% mAP, representing the lowest performance among visual-based mapping methods. The limited performance stems from its reliance on post-processing through clustering and Non-Maximum Suppression (NMS) to generate vectorized map elements, which introduces errors and inefficiencies.

InstaGraM [37] significantly improves upon the CNN-based paradigm by introducing E2E decoding strategies that reduce post-processing requirements. The method combines CNNs and graph neural networks to extract and relate map elements, using two CNNs to detect vertices and edges, followed by an attentional GNN for vertex association. This architectural improvement yields 36.7% mAP, representing a substantial 13.7% improvement over

HDMapNet [19], demonstrating the importance of E2E learning in mapping tasks.

Transformer-Based Mapping Methods

Transformer-based methods have revolutionized visual-based mapping by leveraging self-attention mechanisms to capture long-range dependencies crucial for accurate map construction. These methods demonstrate consistently superior performance compared to CNN-based approaches, with all Transformer methods achieving above 48% mAP.

MapTR [23] pioneered single-stage parallel decoding in mapping, addressing efficiency bottlenecks of autoregressive approaches. It initializes hierarchical queries with instance-level and point-level embeddings, achieving 50.3% mAP. InsMapper [51] (48.3% mAP) focuses on leveraging inner-instance information for improved detection, while MapVR [55] (51.2% mAP) introduces rasterization supervision for enhanced geometric understanding.

Several methods achieve significant performance improvements through sophisticated representation designs. PivotNet [11] (57.6% mAP) introduces pivot-based vectorized representation selecting key geometric points, demonstrating a 7.3% improvement over MapTR [23]. BeMapNet [33] (59.8% mAP) employs piecewise Bezier curves to capture complex map element shapes, achieving the highest performance among curve-based representations.

The most substantial improvements come from advanced query and decoder designs. MapTRv2 [24] (61.5% mAP) introduces one-to-many matching mechanisms and auxiliary dense supervision, improving upon the original MapTR by 11.2%. MGMap [25] (64.8% mAP) leverages learned masks for enhanced localization, integrating global structural information from instance masks. MapQR [27] (66.4% mAP) proposes a "scattering and aggregation" mechanism that distributes queries into specialized sub-queries, achieving competitive performance.

HIMap [61] represents the pinnacle of query-based approaches, achieving 66.7% mAP through HIQuery-a hybrid representation integrating point-level and element-level information via point-element interaction modules. However, the most significant breakthrough comes from mask-based approaches: Mask2Map [10] (71.6% mAP) and MGMapNet [52] (73.6% mAP) demonstrate that mask-guided learning provides substantial advantages for mapping tasks. MGMapNet's [52] 73.6% mAP represents a 50.6% relative improvement over HDMapNet [19], highlighting the evolution of visual-based mapping capabilities.

Performance Analysis

Table 3.3 presents the comprehensive performance comparison of visual-based HD mapping methods without prior information, revealing critical insights about architectural effectiveness and evolutionary trends.

Several key performance insights emerge from this analysis. First, Transformer-based methods demonstrate overwhelming superiority over CNN-based approaches. The performance gap is substantial: the best CNN method (InstaGraM, 36.7% mAP) is outperformed by even the weakest Transformer method (InsMapper [51], 48.3% mAP) by 11.6%. This gap widens dramatically when comparing against state-of-the-art Transformer methods, with

TABLE 3.3: Performance of HD Mapping Methods Without Prior Information

Method	Conference	Backbone	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP	Key Innovation
HDMapNet [19]	ICRA'21	CNN	14.4	21.7	33.0	23.0	FCN with post-processing
InstaGraM [37]	T-ITS'25	CNN	40.8	30.0	39.2	36.7	CNN+GNN association
InsMapper [51]	2024	Transformer	44.4	53.4	52.8	48.3	Instance-based mapping
MapTR [23]	2023	Transformer	46.3	51.5	53.1	50.3	Single-stage parallel decoding
MapVR [55]	2023	Transformer	47.7	54.4	51.4	51.2	Virtual reality approach
PivotNet [11]	ICCV'23	Transformer	56.5	56.2	60.1	57.6	Pivot-based representation
BeMapNet [33]	CVPR'23	Transformer	57.7	62.3	59.4	59.8	Bezier curve representation
MapTRv2 [24]	IJCV'24	Transformer	59.8	62.4	62.4	61.5	Improved MapTR version
MGMap [25]	CVPR'24	Transformer	61.8	65.0	67.5	64.8	Multi-granularity approach
MapQR [27]	ECCV'24	Transformer	68.0	63.4	67.7	66.4	Query-based representation
HIMap [61]	CVPR'24	Transformer	62.6	68.4	69.1	66.7	Hierarchical approach
Mask2Map [10]	ECCV'24	Transformer	70.6	71.3	72.9	71.6	Mask-based approach
MGMapNet [52]	ICLR'25	Transformer	71.8	74.3	74.8	73.6	Multi-granularity network

MGMapNet (73.6% mAP) achieving double the performance of InstaGraM.

Second, clear performance progression exists within Transformer methods. Early approaches (MapTR [23]: 50.3%, MapVR: 51.2%) establish baseline capabilities, while advanced representation learning methods (PivotNet: 57.6%, BeMapNet: 59.8%) demonstrate steady improvements. The latest mask-based approaches (Mask2Map: 71.6%, MGMapNet: 73.6%) represent breakthrough performance levels.

Third, different map elements show varying detection difficulties. Boundary detection ($AP_{bou.}$) generally achieves the highest scores across methods, while pedestrian crossing detection ($AP_{ped.}$) often represents the most challenging task. MGMapNet demonstrates balanced performance across all element types, indicating robust general-purpose mapping capabilities.

Fourth, methods published in 2024-2025 show significant performance leaps. The transition from query-based approaches (HIMap [61]: 66.7%, MapQR: 66.4%) to mask-based methods (Mask2Map: 71.6%, MGMapNet: 73.6%) represents a paradigm shift, with mask guidance providing 5-7% mAP improvements over purely attention-based approaches.

Alternative Dataset Approaches

Several noteworthy methods employ different datasets and evaluation protocols, limiting direct performance comparison with the nuScenes-based methods discussed above. LaneSegNet [20] proposes Lane Segment perception as a novel map learning formulation, introducing an E2E network specifically designed for this task. The method incorporates two significant innovations: a lane attention module employing a head-to-region mechanism to capture long-distance attention relationships, and a reference point initialization strategy that enhances positional prior learning for lane attention mechanisms. Experimental results on the OpenLane-V2 dataset demonstrate state-of-the-art performance, validating the effectiveness of the proposed lane segment formulation. However, due to the fundamental differences in dataset characteristics and evaluation metrics between OpenLane-V2 and nuScenes, direct performance comparison with the methods analyzed in this survey presents limited comparability and therefore warrants separate consideration in future comprehensive evaluations.

3.2.2 Temporal Prior Information Methods

Temporal prior methods leverage short-term information from consecutive frames to improve mapping consistency and accuracy. These approaches address temporal instability in frame-by-frame mapping while maintaining computational efficiency for real-time applications. The integration of temporal information demonstrates substantial performance improvements over no-prior baselines, with the best temporal methods achieving over 73% mAP.

Streaming and Memory-Based Approaches

StreamMapNet [54] pioneered streaming temporal fusion with dual strategies: query propagation retaining high-confidence element queries across frames, and BEV fusion aligning and merging features from consecutive frames. Despite its foundational importance, StreamMapNet achieves 63.4% mAP, representing moderate performance among temporal methods. The relatively lower performance suggests that simple temporal fusion strategies, while beneficial, require more sophisticated designs to fully exploit temporal information.

SQD-MapNet [42] introduces stream query denoising mechanisms to enhance temporal consistency, adding controlled noise to previous frame elements and recovering geometric shapes through denoising. This approach achieves 63.9% mAP, showing minimal improvement over StreamMapNet (0.5% mAP), indicating that denoising alone provides limited benefits for temporal fusion.

MapTracker [9] adopts a more sophisticated dual-memory mechanism, with BEV memory modules selecting relevant features from historical frames based on geometric distance, while vector memory modules filter historical map element queries. This architectural advancement yields 71.9% mAP, representing an 8.5% improvement over StreamMapNet and demonstrating the effectiveness of selective memory utilization.

Historical and Predictive Approaches

HRMapNet [56] utilizes historical rasterized maps for enhanced HD mapping, employing feature aggregation modules to fuse current BEV features with rasterized map features. The method achieves 67.2% mAP, showing that historical map integration provides meaningful improvements but falls short of the best temporal approaches.

PrevPredMap [31] integrates high-level information from previous predictions, including map element categories, confidence scores, and spatial locations. This method achieves 67.6% mAP, slightly outperforming HRMapNet by 0.4%, suggesting that prediction-level temporal fusion provides comparable benefits to feature-level approaches.

PriorMapNet [41] leverages prior map information for enhanced performance, achieving 67.1% mAP. The similar performance to PrevPredMap (67.6% mAP) indicates that different temporal integration strategies converge to comparable effectiveness levels when operating at similar sophistication levels.

Advanced Temporal Integration

MapUnveiler [17] employs temporal modeling for map unveiling, achieving 68.0% mAP. While this represents solid performance, it demonstrates that temporal modeling alone, without sophisticated integration mechanisms, provides limited advantages over simpler approaches.

HisTrackMap [53] represents a significant breakthrough in temporal prior utilization, achieving 73.8% mAP through advanced historical tracking mechanisms. This 10.4% improvement over StreamMapNet demonstrates that sophisticated temporal tracking and historical information integration can provide substantial performance gains.

The most impressive advancement comes from Uni-PrevPredMap [30], which achieves 74.0% mAP using temporal information alone. This represents the highest performance among pure temporal methods and demonstrates a remarkable 10.6% improvement over StreamMapNet, indicating that unified temporal prediction frameworks provide superior temporal information utilization.

Performance Analysis

Table 3.4 demonstrates the effectiveness of temporal prior utilization, showing substantial improvements over no-prior baselines and revealing the relative effectiveness of different temporal integration strategies.

TABLE 3.4: Performance Comparison of Temporal Prior Methods

Method	Conference	$AP_{ped.}$	$AP_{div.}$	$AP_{bou.}$	mAP	Temporal Strategy
StreamMapNet [54]	WACV'24	61.9	66.3	62.1	63.4	Streaming fusion
SQD-MapNet [42]	ECCV'24	63.0	62.5	63.3	63.9	Query denoising
PriorMapNet [41]	2024	64.0	69.0	68.2	67.1	Prior integration
PrevPredMap [31]	WACV'25	66.3	66.9	64.5	67.6	Prediction fusion
MapUnveiler [17]	NeurIPS'24	67.6	67.6	68.8	68.0	Temporal modeling
MapTracker [9]	ECCV'24	75.3	69.2	71.2	71.9	Dual-memory
HisTrackMap [53]	2025	76.9	72.7	71.9	73.8	Historical tracking
Uni-PrevPredMap [31]	2025	76.2	72.3	73.6	74.0	Unified temporal

Several key performance insights emerge from this analysis. First, temporal methods clearly stratify into three performance tiers. Basic methods (StreamMapNet, SQD-MapNet) achieve 63-64% mAP, intermediate methods (PriorMapNet, PrevPredMap, MapUnveiler) reach 67-68% mAP, while advanced methods (MapTracker, HisTrackMap, Uni-PrevPredMap) exceed 71% mAP.

Second, the performance gap between basic and advanced temporal methods (10+ mAP points) exceeds the improvement from no-prior to basic temporal methods (approximately 7-8 mAP points), indicating that temporal integration sophistication matters more than simple temporal information inclusion.

Third, methods emphasizing tracking mechanisms (HisTrackMap: 73.8%, MapTracker: 71.9%) outperform those focusing primarily on feature fusion(StreamMapNet: 63.4%), suggesting that explicit temporal correspondence tracking provides superior benefits to implicit feature-level fusion.

Fourth, Uni-PrevPredMap's [30] 74.0% mAP represents superior performance achieved

through balanced integration of temporal and historical priors, specifically using a 50% non-prior and 50% temporal-prior mixing strategy. This result demonstrates that moderate incorporation of historical information achieves optimal model performance, as the unified framework can adaptively leverage both temporal perception continuity and offline map constraints rather than relying exclusively on any single prior source.

3.2.3 Long-term Historical Prior Methods

Long-term prior methods represent the most sophisticated approach to HD map construction, incorporating historical map information spanning extended temporal periods. These methods demonstrate the highest performance gains by leveraging accumulated knowledge from previous traversals and global map priors.

Neural Map Prior Approaches

Neural Map Prior (NMP) [50] introduced the concept of global neural map priors, employing cross-attention mechanisms to integrate BEV features with accumulated global prior features. The method utilizes a gated recurrent unit (GRU) to dynamically update global map representations, enabling continuous learning from traversal history.

The neural map prior framework addresses fundamental limitations of single-frame approaches by maintaining persistent representations of road infrastructure. The cross-attention integration allows selective incorporation of relevant prior knowledge while adapting to environmental changes, demonstrating robustness to seasonal variations and temporary modifications.

Historical Rasterized Map Integration

HRMapNet [56] pioneered the utilization of historical rasterized maps for enhanced map element detection. The method employs a feature aggregation module to fuse current BEV features with rasterized map features, enriching representation for improved detection accuracy. Vectorized map predictions are subsequently rasterized and integrated into a global map framework. Similar to P-MapNet's integration strategy, HRMapNet requires combination with base architectures to achieve performance improvements. When integrated with MapTRv2, the framework demonstrates the effectiveness of historical rasterized map information for long-term prior utilization, providing substantial performance enhancements through persistent map representation maintenance.

SD Map Integration

P-MapNet [16] demonstrates effective integration of Standard Definition (SD) maps as complementary prior information. The method rasterizes SD maps, encodes them with CNNs, and adaptively fuses SD map features with BEV features. This approach leverages readily available SD map data to provide structural constraints and topological guidance for HD map

construction, proving particularly beneficial for distant map element detection and structural correction.

Outdated HD Map Integration

A distinct category of long-term prior methods represents a paradigmatic shift from traditional HD map generation to HD map updating approaches. These methods leverage outdated or imperfect existing HD map information to enhance online map construction, fundamentally changing the problem formulation from creating maps de novo to correcting and updating existing cartographic resources.

Traditional HD map generation approaches create vectorized maps entirely from sensor observations without prior cartographic knowledge. In contrast, HD map updating methodologies assume the availability of existing map representations that require correction or temporal synchronization with current environmental conditions. This distinction carries significant practical implications for autonomous driving deployment, as updating scenarios reflect real-world operational contexts where some form of prior mapping infrastructure typically exists. MapEX Framework for Map Updating "Mind the Map" [39] introduces MapEX as a novel query-based framework specifically designed for HD map updating rather than pure generation. The method addresses practical scenarios where existing maps contain temporal inconsistencies or localization errors, identifying three realistic updating scenarios: minimalist maps requiring element completion, noisy maps requiring spatial correction, and outdated maps requiring temporal synchronization. MapEX's technical architecture reflects this updating paradigm through two key innovations: non-learnable existing (EX) queries that encode prior map elements as initialization points, and a pre-attribution mechanism that maintains correspondence tracking between existing and updated map elements. This approach fundamentally differs from generation methods by providing structured prior knowledge that guides the updating process rather than learning spatial relationships entirely from sensor observations.

This method demonstrates substantial performance improvements in map updating scenarios on the nuScenes dataset. The most significant performance occurs in temporal evolution scenarios where existing HD maps have become outdated due to real-world changes. In Scenario 3a, which simulates substantial environmental modifications by deleting 50% of pedestrian crossings and lane dividers while adding new elements and applying warping distortions, MapEX achieves 85.9% mAP. Scenario 3b accounts for the realistic situation where substantial portions of maps remain unchanged over time, randomly mixing true HD maps with perturbed versions (p=0.5), achieving an impressive 93.1% mAP. [39] These scenarios reflect common real-world situations where painted markers are displaced, intersections are remodeled, or districts are renovated, but HD maps are only updated every few years to reduce maintenance costs.

These results establish MapEX as the first method to systematically address HD map updating by directly integrating existing maps corresponding to specific sensor locations. Rather than generating maps from sensor data alone, the framework leverages sensor observations to correct outdated existing maps, demonstrating that updating paradigms can achieve su-

perior performance while reducing computational requirements compared to pure generation approaches. This updating-focused approach carries important implications for scalable autonomous driving deployment, as it enables efficient map maintenance workflows that build upon existing cartographic infrastructure rather than requiring complete map reconstruction for every operational environment.

Uni-PrevPredMap [30] extends this paradigm by establishing a unified prior-informed framework that systematically integrates temporal perception buffers with simulated outdated HD maps, representing a methodological advancement beyond single-prior approaches. The framework addresses a fundamental limitation in existing methods by recognizing that temporal perception buffers and cost-efficient alternative maps inherently form complementary prior sources for online vectorized HD map construction.

The framework introduces two core technological innovations that distinguish it from existing approaches. The tile-indexed 3D vectorized global map processor enables efficient 3D prior data refreshment, storage, and retrieval through geolocation-synchronized tile partitioning, eliminating complex post-processing operations while maintaining real-time performance. Unlike MapEX's query-based integration approach, this processor implements dual-axis geospatial indexing where each tile contains discrete map vectors confined to respective geographical boundaries. The tri-mode operational optimization paradigm ensures operational robustness across three critical scenarios: non-prior initialization, temporal-prior operation, and temporal-map-fusion-prior navigation. This systematic approach preserves consistency across different operational modes while reducing dependence on idealized map fidelity assumptions through balanced integration strategies.

Uni-PrevPredMap [30] introduces three distinct simulation methodologies in Fig 3.1 that model realistic scenarios where map information becomes temporally inconsistent, representing a more systematic approach compared to MapEX's scenario-based modifications. Minor infrastructure modification applies stochastic displacement magnitudes to statistically 50% of map vectors with a mean 3-meter displacement, simulating common real-world infrastructure changes including divider modifications, boundary adjustments, and pedestrian crossing relocations.

Major infrastructure renovation extends this approach by applying distinct displacement magnitudes to all map vectors per frame, representing extreme boundary conditions for evaluating framework robustness under comprehensive map degradation. Pose misalignment applies identical displacement magnitudes to all map vectors, simulating systematic positioning errors that occur in real-world operations due to localization drift or calibration issues.

The fundamental distinction between Uni-PrevPredMap and MapEX lies in their integration philosophies and architectural implementations. MapEX focuses on direct existing map integration through non-learnable EX queries and pre-attribution mechanisms, essentially treating the problem as map updating rather than unified prior fusion. Uni-PrevPredMap addresses the broader challenge of complementary prior source integration, systematically combining temporal perception continuity with offline map constraints.

While MapEX demonstrates effectiveness in specific updating scenarios with predetermined correspondence tracking, Uni-PrevPredMap achieves superior performance through

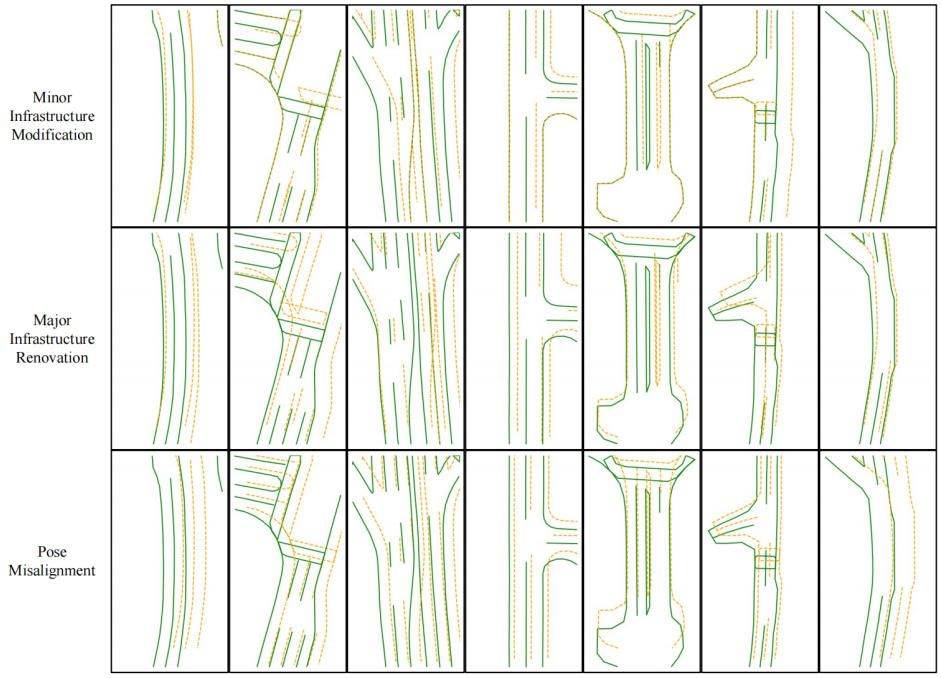

FIGURE 3.1: Comparison between simulated outdated HD maps (orange dashed lines) and ground truth annotations (green solid lines). [30]

synergistic complementarity between temporal observations and simulated outdated maps. The framework's tri-mode optimization strategy enables operational flexibility that MapEX's scenario-specific approach cannot match, particularly in environments with varying prior information availability.

Experimental validation demonstrates the substantial benefits of unified prior integration compared to single-source approaches. The framework achieves 74.0% mAP in temporal-only mode and 80.9% mAP when combining temporal and map priors, representing significant improvements over baseline performance of 64.9% mAP in non-prior scenarios. This performance progression empirically confirms the complementary nature of temporal buffers and simulated outdated maps in enhancing perception stability. The systematic evaluation of mixing ratios reveals optimal integration strategies, with a 0.50:0.30:0.20 ratio for non-prior, temporal-prior, and temporal-map-fusion-prior modes achieving balanced performance across different operational scenarios. Performance adaptively adjusts to displacement magnitude variations, demonstrating framework robustness under varying map degradation conditions while maintaining superior performance compared to temporal-only configurations. The superior performance of these unified integration methods stems from their systematic exploitation of complementary information sources rather than reliance on single-modal prior knowledge. Unlike approaches that depend exclusively on neural map priors or temporal features, Uni-PrevPredMap leverages the inherent complementarity between backward temporal continuity and forward-looking topological constraints, establishing a methodological foundation for robust online HD map construction in dynamic operational environments.

Moreover, recent breakthrough implementations, particularly the RTMap framework [12], demonstrate how simultaneous localization, mapping, and change detection can be integrated within a single E2E architecture. RTMap [12] introduces a paradigmatic approach that

processes multi-sensor observations through hybrid query mechanisms, enabling the system to simultaneously detect matched map elements, identify outdated information requiring removal, and incorporate newly observed features. This unified processing eliminates the temporal delays and error accumulation characteristic of sequential pipelines while enabling real-time performance suitable for autonomous driving applications.

Performance Analysis

Table 3.5 quantifies the significant performance improvements achieved through long-term prior integration, based on comprehensive experimental results.

TABLE 3.5: Performance Impact of Long-term Historical Priors

Method	Prior Type	w/o	w/	Δ	Integration Strategy
HDMapNet [19]	SD+HD	27.7	30.7	+3.1	P-MapNet [16]
VectorMapNet [26]	HD	40.9	44.8	+3.9	Neural Map Prior [50]
StreamMapNet [54]	HD	60.4	66.3	+5.9	HRMapNet [56]
MapTRv2 [24]	HD	61.5	67.2	+5.7	HRMapNet [56]

3.2.4 Discussions

The comprehensive analysis of prior information utilization reveals fundamental insights into the evolution and capabilities of HD map construction systems. The progression from no-prior methods to sophisticated historical prior integration demonstrates a clear trajectory of performance enhancement and system maturity.

Quantitative Performance Analysis

The performance improvements across different prior information categories demonstrate a consistent trend: methods incorporating long-term historical priors achieve the most significant gains, with improvements ranging from 3.0% to 5.9% mAP. The data from Table 3.5 shows that historical rasterized map integration (StreamMapNet [54] and MapTRv2 [24]) provides the largest improvements of approximately 6% mAP, while SD map integration through P-MapNet [16] offers more modest but still substantial gains of 3.0-3.9% mAP.

Comparing architectural approaches, Transformer-based methods without priors achieve 48.3-66.7% mAP (Table 3.3), while the same architectures with long-term priors reach 66.3-67.2% mAP (Table 3.5), representing a ceiling effect where sophisticated priors become essential for achieving state-of-the-art performance.

Architectural Implications

The integration of prior information fundamentally alters architectural requirements. No-prior methods require sophisticated view transformation and feature extraction capabilities, with Transformer architectures proving superior due to their ability to capture long-range dependencies essential for map element detection.

Temporal prior methods introduce memory management challenges but demonstrate that even simple temporal fusion strategies can provide meaningful improvements in detection consistency. The dual-memory approaches in MapTracker and streaming mechanisms in StreamMapNet represent efficient solutions balancing performance gains with computational overhead.

Long-term prior methods achieve the highest performance ceiling but require sophisticated data management infrastructures. The neural map prior approach (NMP) [50] and historical rasterized map integration (HRMapNet [56]) demonstrate that persistent map representations can provide substantial performance benefits, justifying the additional architectural complexity. The unified framework introduced by Uni-PrevPredMap [30] represents a significant advancement in long-term prior utilization, achieving state-of-the-art performance through strategic integration of temporal perception buffers and cost-efficient alternative maps, demonstrating the complementary nature of different prior information sources.

Research Directions

The consistent performance improvements from prior information integration suggest several promising research directions. Hybrid approaches combining multiple types of prior information could yield further gains. The significant improvements from historical rasterized maps (5.7-5.9% mAP) indicate that persistent map representations deserve continued investigation.

The development of efficient prior information sharing mechanisms could enable collaborative mapping systems where vehicles contribute to and benefit from shared historical knowledge. Additionally, adaptive prior selection based on environmental context and learned prior representations offer potential for continued advancement in mapping accuracy and reliability.

Chapter 4

Experiments

The fundamental challenge in deploying learning-based HD mapping and localization systems lies in the significant discrepancy between idealized experimental conditions and real-world operational uncertainties. Contemporary research predominantly evaluates these systems under conservative perturbation assumptions, typically limiting initial pose uncertainties to 2-5 meters and rotational errors to a few degrees. However, this experimental paradigm inadequately reflects the harsh realities of autonomous vehicle deployment, where GPS signal degradation in urban canyons, tunnels, and dense urban environments can introduce positional uncertainties exceeding 30 meters and heading errors reaching 10°or more. This disparity between laboratory validation and deployment scenarios creates a critical knowledge gap that undermines the reliability of current systems when transitioning from controlled experimental settings to real-world operation. The conservative perturbation assumptions employed in existing literature not only fail to capture realistic sensor uncertainty characteristics but also risk introducing data leakage artifacts where models inadvertently learn dataset-specific patterns rather than robust spatial reasoning capabilities.

To address these fundamental limitations, this chapter investigates two critical research questions that directly impact the practical deployment of integrated mapping-localization systems:

Research Question 1: How does initial pose uncertainty affect map-based localization robustness? Current localization methods demonstrate excellent performance under minimal perturbations but their behavior under realistic GNSS degradation scenarios remains largely uncharacterized. Understanding this relationship is essential for designing robust systems that maintain operational safety across varying environmental conditions.

Research Question 2: How do pose perturbations impact HD mapping system performance? While many HD mapping approaches incorporate ablation studies demonstrating resilience to small pose errors, the impact of realistic localization uncertainties on mapping quality requires systematic investigation. This understanding is crucial for both HD map generation and updating scenarios where precise ego-pose estimation directly affects prior map utilization effectiveness.

Our experimental design addresses these questions through carefully controlled perturbation studies that avoid data leakage risks while simulating realistic deployment conditions.

The investigation spans both fine-grained uncertainties representative of high-precision positioning systems and large-scale errors characteristic of challenging urban environments. By systematically varying perturbation magnitudes during both training and testing phases, we establish empirical foundations for optimal system design parameters that balance robustness with accuracy requirements.

The experimental findings provide critical insights for architectural design decisions in integrated mapping-localization frameworks, establishing performance thresholds and robustness characteristics essential for safe autonomous vehicle deployment. These results directly inform the unified framework proposed in Chapter 5, ensuring that theoretical architectural advances translate effectively to practical implementation requirements.

4.1 Map Based Localization

FIGURE 4.1: Comparison of localization robustness under different initial pose perturbation scenarios: (a) no perturbation, (b) XY translation perturbation, (c) orientation perturbation, (d) combined perturbation (translation + orientation).

4.1.1 Experimental Design and Motivation

The fundamental challenge in visual localization systems lies in the inherent uncertainty of sensor measurements in real-world deployment scenarios. GPS receivers typically exhibit accuracy limitations of 3-20 meters in urban environments, while IMU systems suffer from drift and calibration errors that compound over time [6]. This discrepancy between idealized training conditions and practical deployment environments creates a critical domain gap that significantly impacts system performance. Fig 4.1 shows how to apply combined perturbation (translation + orientation) to the GT pose.

To address this challenge systematically, we design a comprehensive ablation study that investigates the impact of initial pose perturbations during both training and testing phases on the robustness of visual localization systems. The experimental hypothesis posits that models trained with realistic sensor noise will demonstrate superior performance under actual deployment conditions compared to models trained with perfect initial pose estimates.

The experimental design follows three core principles. First, we isolate the effects of translational and rotational uncertainties through carefully designed perturbation configurations. Second, we ensure fair comparison by maintaining identical training procedures across all configurations, varying only the perturbation parameters. Third, we evaluate all trained models under multiple test conditions to comprehensively assess their practical applicability across different deployment scenarios.

4.1.2 Experimental Setup

Dataset and Model Configuration

We conduct experiments using the KITTI dataset, a widely-adopted benchmark for autonomous driving research that provides high-quality camera imagery with precise ground truth poses obtained from Real-Time Kinematic (RTK) GPS systems. The images are captured by cameras mounted on a vehicle driving through urban and residential areas, providing about $\pm$20m and $\pm 10\circ$ of the GT position accuracy through RTK [36].

The KITTI dataset contains stereo images captured by a moving vehicle from different trajectories at different times, with minimal trajectory revisitation. We split the entire raw dataset into three subsets: Training, Test1, and Test2. The Training and Test1 sets originate from the same geographical region, while Test2 represents a different area to evaluate generalization capability. Training is performed on the Training set, with Test1 used for validation during training, while trajectories within 5 meters of validation data are removed from the training set to prevent overfitting. The dataset distribution comprises 19,655 training images, 3,773 Test1 images, and 7,542 Test2 images.

Our approach builds upon OrienterNet [35], a state-of-the-art visual localization model pre-trained on the large-scale MGL dataset. We fine-tune this pre-trained model on KITTI using different initial pose perturbation strategies to assess their impact on localization robustness.

Perturbation Configurations

We evaluate five distinct perturbation configurations designed to isolate different aspects of sensor uncertainty:

- **Configuration 1 (0m, 0°):** Baseline configuration with no perturbations, representing idealized sensor conditions
- **Configuration 2 (10m, 5°):** Moderate perturbations simulating typical GPS and IMU accuracy in favorable conditions
- **Configuration 3 (20m, 0°):** Pure translational perturbations to isolate the impact of GPS positional uncertainty
- **Configuration 4 (20m, 10°):** Combined perturbations representing realistic sensor conditions in challenging environments, corresponding to the original OrienterNet setup
- **Configuration 5 (30m, 10°):** High perturbations simulating degraded sensor performance in adverse conditions

During training, perturbations are applied to the ground truth poses using pre-computed shift values sampled from uniform distributions within the specified radius for both translational and angular components. This ensures reproducibility across experiments while maintaining realistic noise characteristics. For testing, we apply the same five perturbation configurations to create a comprehensive 5×5 experimental matrix totaling 25 distinct train-test combinations.

Training Protocol

All experiments utilize identical training hyperparameters to ensure fair comparison. We employ a learning rate of 0.0001 and distribute training across three NVIDIA A40 GPUs. For each configuration, we monitor validation performance and select the best epoch before overfitting occurs [35]. Configurations with excessive perturbations that fail to converge are excluded from analysis to focus on practically viable training strategies.

To comprehensively assess model robustness, we evaluate each trained model under all five test perturbation configurations. This systematic approach enables us to understand how training perturbations affect performance across different deployment scenarios, from ideal conditions to challenging real-world environments.

4.1.3 Results and Analysis

Performance Pattern Analysis

Figure 4.2 presents a comprehensive analysis of system performance across translational and angular perturbation dimensions, revealing critical insights for robust visual localization system design. The experimental results demonstrate fundamental differences between

perturbation-aware and traditional training approaches under realistic deployment conditions.

The naming convention Kitti_X_Y represents training configurations where uniform distribution perturbations are applied to initial poses during training, with X indicating translational perturbation range (in meters) and Y indicating angular perturbation range (in degrees). For example, Kitti_20_10 indicates training with up to 20m translational and 10°angular perturbations applied to initial poses. The test conditions (e.g., Pure Trans (20m, 0°)) refer to perturbations applied during inference to evaluate model robustness.

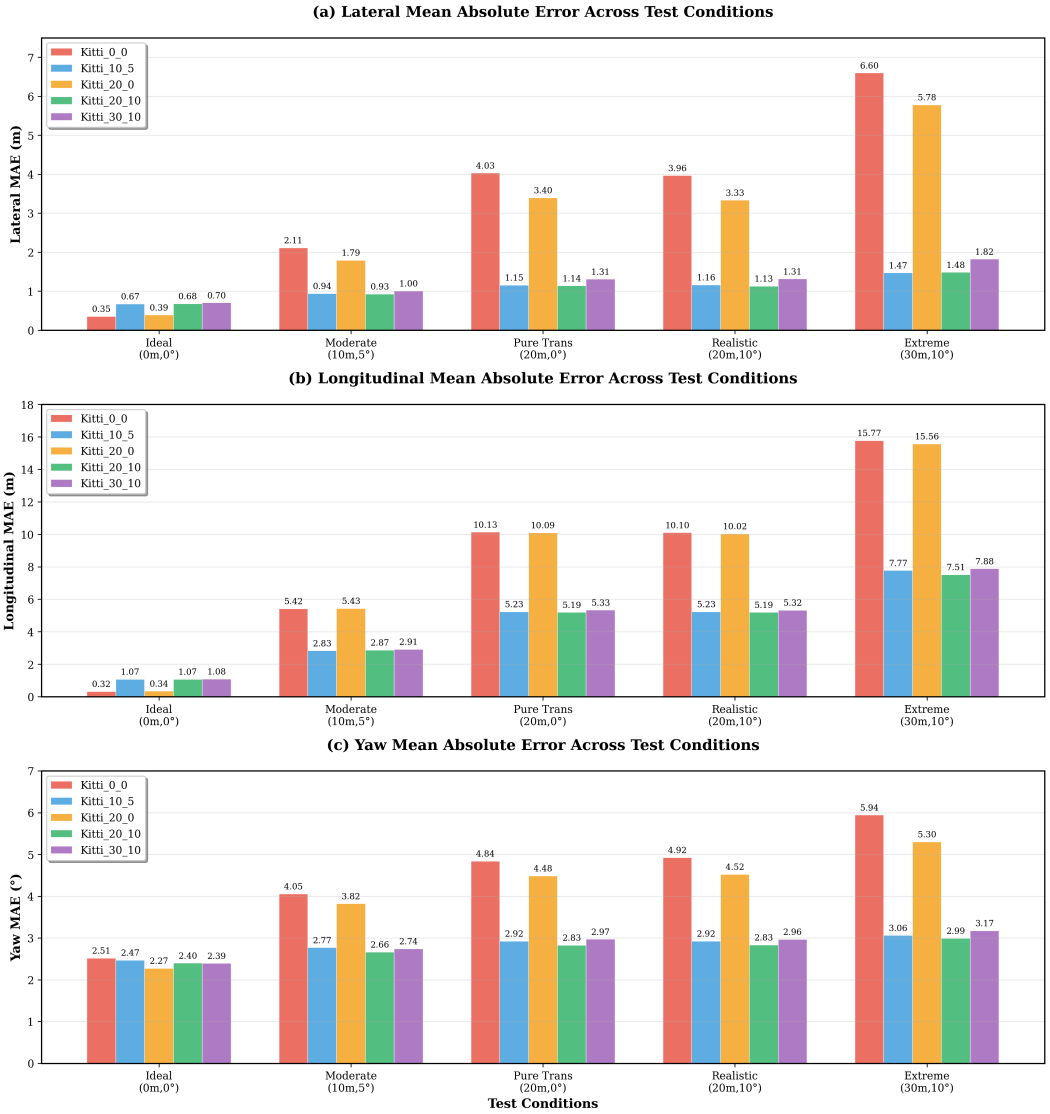

FIGURE 4.2: Visualization results on KITTI dataset test perturbation

Lateral Performance Analysis Across Training Configurations: The lateral MAE results in Figure 4.2(a) reveal distinct performance patterns across the five test conditions. Under ideal conditions (0m, 0°), all configurations achieve comparable performance, with Kitti_0_0 (baseline) achieving the lowest lateral MAE of 0.35m. However, as test conditions become more challenging, dramatic performance divergences emerge. The Kitti_0_0 configuration exhibits severe performance degradation, reaching 6.60m lateral MAE under extreme conditions (30m, 10°). In stark contrast, the perturbation-aware training configurations (Kitti_10_5, Kitti_20_10, Kitti_30_10) demonstrate remarkable stability, maintaining lat-

eral MAE consistently below 1.5m across all test conditions. Notably, Kitti_10_5 achieves the best overall lateral performance with 0.94m under moderate conditions and only 1.47m under extreme conditions, representing a 4.5-fold improvement over baseline performance under challenging scenarios.

Longitudinal Estimation Robustness Under Perturbation Stress: The longitudinal MAE patterns in Figure 4.2(b) demonstrate even more pronounced performance differences between training approaches. The Kitti_0_0 configuration experiences catastrophic performance degradation from 0.32m under ideal conditions to 15.77m under extreme conditions, representing nearly a 50-fold increase in error. Pure translational training (Kitti_20_0) exhibits similar vulnerability with comparable degradation patterns, reaching 15.56m under extreme conditions. Conversely, perturbation-aware configurations maintain controlled longitudinal errors across all test conditions. Kitti_20_10 demonstrates the most consistent longitudinal performance, with errors ranging from 1.07m under ideal conditions to 7.51m under extreme conditions. This represents a substantial improvement in robustness, containing the error increase to approximately 7-fold rather than the 50-fold degradation observed in baseline approaches.

Angular Stability Through Mixed Perturbation Training: The yaw MAE analysis in Figure 4.2(c) reveals the superior angular estimation capabilities of perturbation-aware training. All configurations start with comparable yaw performance under ideal conditions (approximately 2.4-2.5°). However, as angular perturbations increase in test conditions, Kitti_0_0 and Kitti_20_0 configurations show progressive degradation, reaching 5.94°and 5.30°respectively under extreme conditions. Perturbation-aware configurations with angular training components demonstrate exceptional angular stability, with Kitti_20_10 maintaining the most consistent performance across all conditions, varying only from 2.40°to 2.99°. This represents less than 0.6°variation compared to the 3.4°variation observed in Kitti_0_0 configuration, indicating superior orientation estimation robustness.

Optimal Configuration Selection Through Comparative Analysis: The comprehensive performance comparison establishes Kitti_20_10 as the optimal training configuration across all estimation dimensions. This configuration achieves the best balance between accuracy preservation and uncertainty tolerance, demonstrating competitive performance under ideal conditions while maintaining superior robustness under challenging scenarios. The systematic comparison reveals that perturbation-aware training approaches consistently outperform both baseline (Kitti_0_0) and pure translational (Kitti_20_0) approaches, with Kitti_20_10 showing the most stable performance profile across lateral, longitudinal, and angular dimensions. The 20m translational and 10°angular perturbation combination in training effectively prepares the system for realistic deployment uncertainties without compromising fundamental estimation accuracy.

The superior performance of Kitti_20_10 can be attributed to several key factors. First, the inclusion of both translational and angular perturbations during training creates a comprehensive uncertainty representation that mirrors real-world deployment scenarios, where sensor noise and calibration errors typically manifest in both spatial and orientational dimensions simultaneously. This joint perturbation strategy enables the network to learn robust feature associations that remain stable under combined uncertainty conditions, unlike

pure translational training (Kitti_20_0) which fails to capture angular uncertainty characteristics. Second, the 20m/10°perturbation magnitude strikes an optimal balance between training challenge and learning stability—sufficient to induce meaningful robustness without introducing excessive noise that could impair convergence or degrade feature learning quality. This is evidenced by comparing Kitti_20_10 with Kitti_30_10, where the increased perturbation magnitude begins to show diminishing returns and occasionally worse performance. Finally, the mixed perturbation approach functions as an effective data augmentation strategy that implicitly regularizes the learning process, forcing the network to develop invariant representations that generalize well across the uncertainty space typically encountered in autonomous vehicle deployment scenarios.

Cross-Condition Stability Analysis

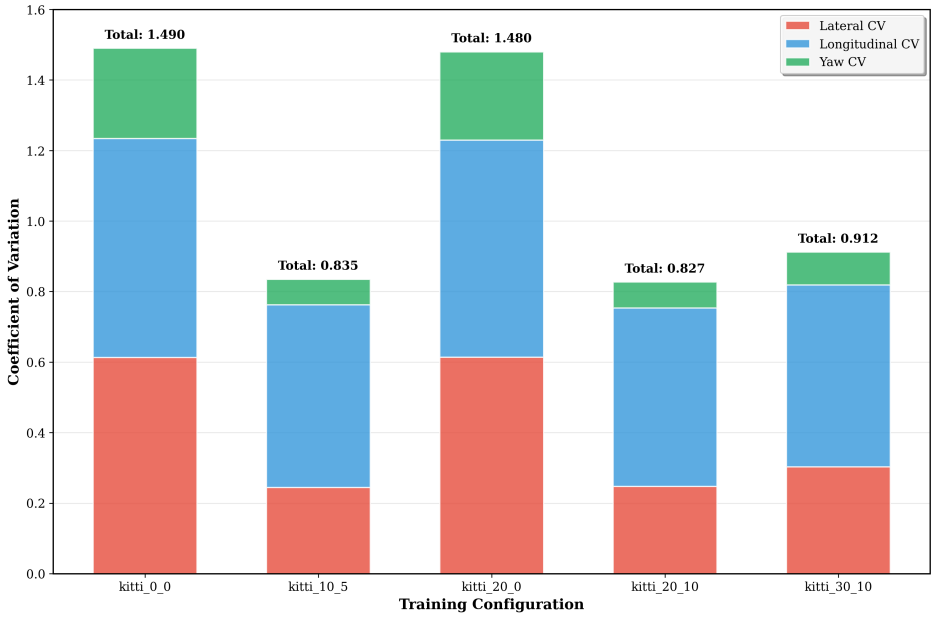

FIGURE 4.3: Coefficient of variation analysis across training configurations, showing dimensional stability contributions for each training configuration.

To quantitatively assess model robustness across deployment scenarios, we conducted comprehensive stability analysis using coefficient of variation (CV) and maximum-to-minimum ratio metrics. The coefficient of variation is calculated as $\text{CV} = \sigma/\mu$, where σ represents the standard deviation and μ the mean across all five test conditions, providing a normalized measure of relative variability independent of absolute performance levels.

Figure 4.3 presents a comprehensive breakdown of coefficient of variation contributions across training configurations, revealing fundamental differences in dimensional stability characteristics. The stacked visualization clearly demonstrates that mixed perturbation training configurations achieve dramatically lower total CV values compared to baseline and pure translational approaches. The kitti_20_10 and kitti_10_5 configurations exhibit total CV values of approximately 0.83 and 0.84 respectively, representing superior stability across all estimation dimensions. This contrasts markedly with baseline (kitti_0_0) and pure translational (kitti_20_0) configurations, which demonstrate total CV values approaching 1.5, indicating substantial performance variability across deployment conditions.

It also reveals that angular estimation achieves exceptional stability in mixed perturbation configurations, with yaw CV contributions remaining minimal across all training approaches. This pattern indicates that orientation estimation responds particularly favorably to combined uncertainty training, developing robust rotational invariance properties that maintain consistent performance regardless of deployment perturbation levels. The longitudinal dimension consistently contributes the largest portion of total CV across all configurations, suggesting inherent challenges in forward-backward estimation that require specialized attention in system design and deployment planning.

The comprehensive stability analysis establishes a clear performance hierarchy that aligns with the MAE-based evaluations presented in Figure 4.2. The kitti_20_10 configuration emerges as the optimal choice, achieving the lowest total CV while maintaining balanced performance across all estimation dimensions. The result demonstrates that mixed perturbation training approaches cluster together with significantly superior stability characteristics compared to traditional methods. The dramatic stability gap between mixed and single-dimension perturbation approaches validates the critical necessity of angular uncertainty inclusion in training protocols, transforming visual localization from brittle high-variance systems to robust deployment-ready solutions suitable for safety-critical autonomous applications.

Visualization Analysis

To comprehensively evaluate the effectiveness of our perturbation-based training approach, we present detailed visualization results across three challenging driving scenarios: bidirectional roads, curved trajectories, and complex intersections. Each visualization includes the camera view, semantic map, likelihood estimation, and neural feature maps, with ground truth poses indicated by red arrows, perturbed input poses by blue arrows, and model predictions by black arrows. The comparison between models trained with 20m/10°perturbations versus baseline models without perturbation augmentation reveals significant performance differences under identical test perturbations.

Bidirectional Road Scenarios

Figure 4.4 presents the localization performance in bidirectional road environments, where symmetric visual features create fundamental challenges for pose estimation. The upper panel demonstrates the robust performance of our perturbation-augmented model when subjected to 20m/10°perturbations during inference. Despite the substantial input noise, the predicted pose (black arrow) closely aligns with the ground truth (red arrow), while the likelihood map exhibits concentrated activation patterns around the correct location. The neural map shows focused attention on relevant road features, indicating that the model has learned to disambiguate between visually similar road segments through robust feature associations.

In contrast, the lower panel reveals significant performance degradation for the baseline model under identical test conditions. The predicted pose exhibits substantial deviation from the ground truth, with the likelihood map showing erroneous activation in regions corresponding to the opposing lane. This failure mode is particularly pronounced in bidirectional road scenarios due to the inherent environmental symmetry. Bidirectional roads present a funda-

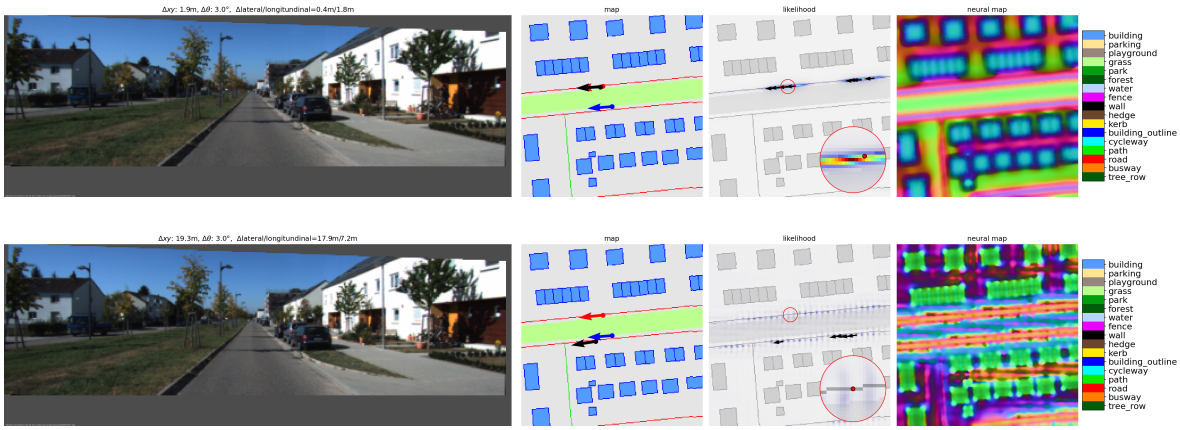

FIGURE 4.4: Visualization results on KITTI dataset test with perturbation on bidirectional road scenario(above: train with perturbation(20m 10°), below: train without perturbation, blue: Init., red: GT, black: Pred.)

mental challenge for visual localization systems due to their near-identical visual appearance on opposing sides. Buildings, vegetation, road markings, and other environmental landmarks often exhibit mirror symmetry across the road centerline, creating what we term "symmetric ambiguity" where perturbed observations may be incorrectly matched to geometrically similar but spatially incorrect map regions.

Our perturbation-based training strategy effectively addresses this challenge by exposing the model to pose uncertainties during training, forcing it to learn more discriminative feature representations that remain invariant to spatial perturbations. The model learns to identify subtle asymmetric cues and develop confidence estimates that prevent spurious matches with opposing lane features, demonstrating enhanced robustness in handling environmental symmetry challenges.

Curved Road Scenarios

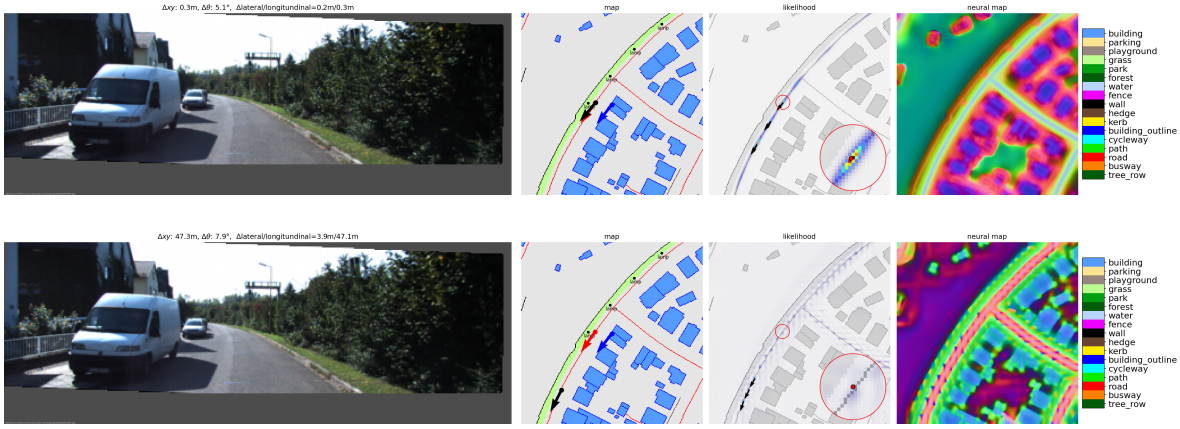

FIGURE 4.5: Visualization results on KITTI dataset test with perturbation on curved road scenario(above: train with perturbation(20m 10°), below: train without perturbation, blue: Init., red: GT, black: Pred.)

Figure 4.5 illustrates the localization performance in curved road environments, which present unique challenges due to their geometric complexity and varying curvature profiles.

The upper panel showcases the remarkable resilience of the perturbation-augmented model, where despite the initial pose estimate (blue arrow) being severely displaced to the roadside, the model's prediction (black arrow) achieves near-perfect alignment with the ground truth (red arrow). The likelihood map exhibits sharp, concentrated activation around the correct location, while the neural feature map demonstrates focused attention on relevant road geometry and landmark features.

Conversely, the lower panel reveals a critical failure mode of the baseline model characterized by longitudinal drift along the curved path. The predicted pose exhibits significant displacement in the direction of the road curvature, with the likelihood map showing diffuse or misaligned activation patterns. This phenomenon occurs because consecutive road segments along curved trajectories exhibit similar cross-sectional profiles but differ in their longitudinal position, creating geometric ambiguity for localization systems.

Curved road environments introduce several fundamental difficulties including geometric ambiguity where points along a curved trajectory may exhibit similar local geometric properties, limited distinctive features in relatively homogeneous environments, perspective distortion from varying viewpoints, and temporal correlation dependency that becomes unreliable under large initial pose errors. Our perturbation-based training strategy effectively addresses these challenges by forcing the model to learn robust feature representations that remain discriminative even under substantial pose uncertainties, encouraging the development of global contextual understanding rather than reliance on local geometric cues alone.

Complex Intersection Scenarios

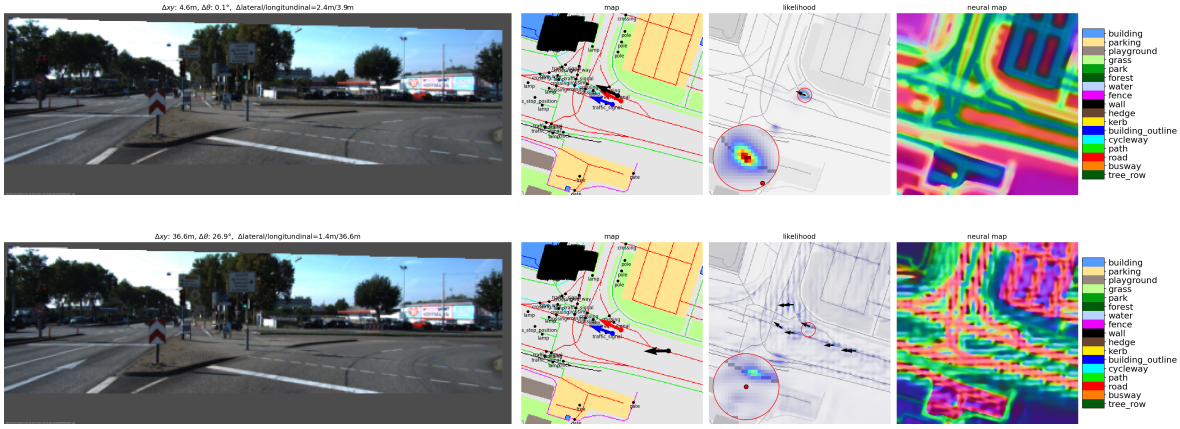

FIGURE 4.6: Visualization results on KITTI dataset test with perturbation on intersection scenario(above: train with perturbation(20m 10°), below: train without perturbation, blue: Init., red: GT, black: Pred.)

Figure 4.6 presents the localization performance in intersection environments, representing one of the most challenging scenarios for autonomous vehicle localization systems. Intersections are characterized by rich semantic information from multiple converging road segments, diverse infrastructure elements, and dynamic occlusions from moving vehicles, creating both opportunities and challenges for accurate pose estimation.

The upper panel demonstrates exceptional robustness of the perturbation-trained model, where despite the initial pose (blue arrow) being displaced to an adjacent lane, the model

successfully recovers the correct position and orientation with near-perfect alignment between predicted and ground truth poses. The likelihood map shows concentrated activation around the true vehicle position, while the neural feature map reveals focused attention on discriminative intersection features such as lane markings, traffic infrastructure, and building facades.

In stark contrast, the lower panel reveals a catastrophic failure of the baseline model, exhibiting both translational and rotational errors. The predicted pose shows significant deviation not only in position but also in heading angle, as evidenced by the misaligned black arrow orientation. The likelihood map displays diffuse or multi-modal activation patterns, suggesting uncertainty in pose estimation across multiple plausible but incorrect locations.

Urban intersections present a unique combination of challenges and opportunities for visual localization systems. While they offer rich semantic context through abundant visual landmarks including traffic signals, road signs, lane markings, and distinctive architectural features, they simultaneously present difficulties through dynamic occlusions from moving vehicles and pedestrians, orientation ambiguity from multiple possible heading directions, scale variations due to open intersection geometry, and complex lighting conditions. The critical angular error observed in the baseline model highlights a fundamental weakness in handling orientation estimation under pose uncertainty, which is particularly crucial in intersection scenarios for distinguishing between different travel directions and ensuring correct lane assignment.

The comprehensive visualization analysis across these three challenging scenarios demonstrates that perturbation-augmented training significantly enhances localization robustness by enabling models to learn discriminative features that remain effective under substantial pose uncertainties. The approach proves particularly valuable in complex geometric environments where traditional localization methods are susceptible to symmetric ambiguities, geometric drift, and orientation errors.

4.2 HD Mapping

4.2.1 Experimental Design and Motivation

The fundamental challenge in online vectorized map perception systems lies in the critical dependence on accurate vehicle localization for historical map maintenance and retrieval. In autonomous driving scenarios, ego-pose estimation systems typically exhibit inherent uncertainties due to GPS accuracy limitations ranging from centimeters to tens of meters in urban environments, combined with IMU drift and calibration errors that accumulate over time. This localization uncertainty directly impacts the core functionality of HRMapNet, where precise ego-pose information is essential for both updating the global historical rasterized map and retrieving relevant local map regions during online perception.

To address this challenge systematically, we design a comprehensive robustness evaluation that investigates two distinct scenarios across different baseline methods: (1) fine-grained localization uncertainties representative of high-precision positioning systems, and (2) large-

scale initial pose errors that may occur during system initialization or under challenging environmental conditions. We evaluate both MapTRv2 and MapQR integrate with HRMapNet which do not apply initial pose error as baseline methods to demonstrate the generalizability and comparative robustness of our design.

4.2.2 Experimental Setup

Dataset and Model Configuration

We conduct experiments using the nuScenes dataset with two state-of-the-art baseline methods: MapTRv2 [24] and MapQR [27], both integrated with the HRMapNet [56] framework. Models are trained for 24 epochs on the standard nuScenes training split and evaluated on the validation set. All robustness evaluations utilize identical pre-trained model weights for each baseline to ensure fair comparison across perturbation configurations.

Two-Scale Perturbation Strategy

We evaluate localization robustness using two complementary experimental designs:

Fine-Scale Perturbations: We inject small-scale Gaussian noise to simulate high-precision localization system uncertainties. Translation noise spans $\sigma_t \in \{0, 0.05, 0.1, 0.2\}$ meters, representing GPS accuracies from perfect to typical urban conditions. Rotation noise covers $\sigma_r \in \{0, 0.005, 0.01, 0.02\}$ radians (approximately $\{0°, 0.3°, 0.6°, 1.1°\}$), reflecting realistic IMU heading uncertainties.

Large-Scale Perturbations: To evaluate system behavior under significant localization failures, we design a second experimental regime with substantially larger noise levels. Translation errors span four levels with standard deviations $\sigma_t \in \{0, 1.67, 3.33, 6.67\}$ meters, corresponding to 99.7% confidence intervals of approximately $\{0, 5, 10, 20\}$ meters. Rotation errors cover $\sigma_r \in \{0, 0.0291, 0.0582\}$ radians (approximately $\{0°, 5°, 10°\}$), simulating compass failures or significant heading estimation errors.

4.2.3 Results and Analysis

Fine-Scale Robustness Performance

TABLE 4.1: MapTRv2+HRMapNet performance under fine-scale localization uncertainties. Results show mAP under varying translation (σ_t) and rotation (σ_r) noise levels.

σ_t (m) / σ_r (rad)	0.000	0.005	0.010	0.020
0.00	67.2	66.7	66.2	64.4
0.05	66.6	66.6	65.9	64.2
0.10	66.7	66.3	65.9	64.2
0.20	66.1	65.3	64.8	63.6

Figure 4.7(a) and (b) present comprehensive heatmap visualizations of fine-scale robustness performance for MapTRv2+HRMapNet and MapQR+HRMapNet, respectively. The

TABLE 4.2: MapQR+HRMapNet performance under fine-scale localization uncertainties. Results show mAP under varying translation (σ_t) and rotation (σ_r) noise levels.

σ_t (m) / σ_r (rad)	0.000	0.005	0.010	0.020
0.00	72.7	72.4	71.9	70.0
0.05	72.7	72.3	71.8	70.0
0.10	72.4	72.3	71.6	69.8
0.20	71.7	71.5	71.0	69.1

visual analysis reveals distinct robustness characteristics between the two baseline methods when subjected to initial pose perturbations that affect both historical map updating and prior map retrieval processes.

The heatmap in Figure 4.7(a) demonstrates MapTRv2+HRMapNet's vulnerability to localization uncertainties, with baseline performance of 67.2 mAP degrading progressively as perturbations increase. As detailed in Table 4.1, the visualization shows clear performance degradation patterns, with the most severe impact occurring under combined translation and rotation noise. At the extreme fine-scale condition ($\sigma_t = 0.2$m, $\sigma_r = 0.02$ rad), performance drops to 63.6 mAP, representing a 3.6-point degradation from baseline. The color gradient clearly illustrates how rotation noise exhibits a more pronounced impact than translation noise, suggesting that heading errors significantly compromise the model's ability to correctly align historical map information during both updating and retrieval operations. In contrast,

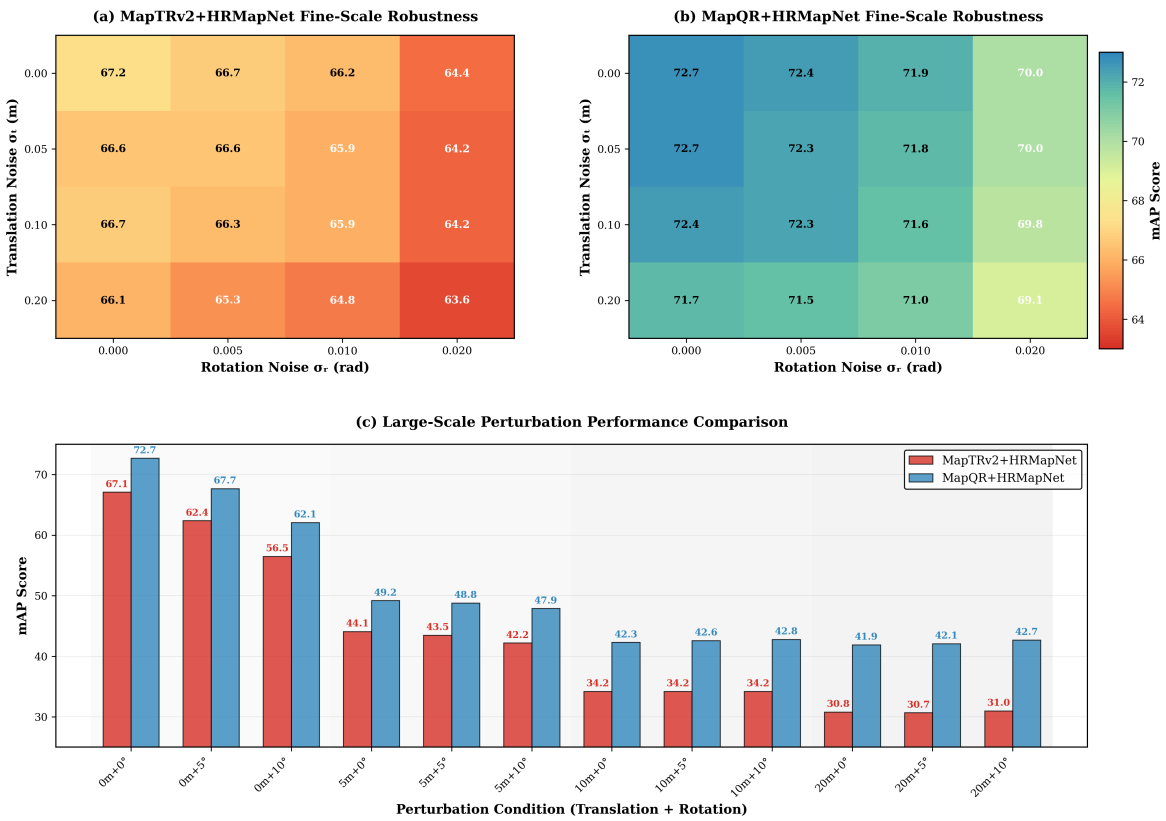

FIGURE 4.7: HRMapNet mapping results on Nuscenes dataset with different test perturbation

Figure 4.7(b) reveals MapQR+HRMapNet's substantially better robustness profile, main-

taining higher performance levels across all perturbation conditions. As shown in Table 4.2, starting from a superior baseline of 72.7 mAP, the system demonstrates remarkable resilience, degrading only to 69.1 mAP under the most challenging fine-scale conditions. The heatmap visualization shows a more gradual color transition compared to MapTRv2+HRMapNet, indicating more stable HD mapping performance when prior map utilization is compromised by pose uncertainties. This suggests that MapQR's query-based architecture maintains more robust spatial reasoning capabilities for historical map integration even when initial pose estimates contain errors.

The side-by-side heatmap comparison reveals that MapQR+HRMapNet maintains a consistent 5.5-point performance advantage over MapTRv2+HRMapNet across all fine-scale perturbation levels. This consistent gap indicates that MapQR's architectural design provides superior resilience to the dual challenges of inaccurate map updating and compromised prior map retrieval that result from pose estimation errors.

TABLE 4.3: MapTRv2+HRMapNet performance under large-scale initial pose errors. Results show individual category performance and overall mAP under substantial localization uncertainties.

Translation / Rotation	**$AP_{div.}$**	**$AP_{ped.}$**	**$AP_{bou.}$**	**mAP**
0m + 0°	67.5	65.4	68.4	67.1
0m + 5°	60.9	62.3	64.0	62.4
0m + 10°	54.4	58.1	56.9	56.5
5m + 0°	44.1	42.4	45.9	44.1
5m + 5°	43.9	41.4	45.2	43.5
5m + 10°	43.4	39.7	43.5	42.2
10m + 0°	35.8	29.0	37.7	34.2
10m + 5°	36.7	27.9	37.9	34.2
10m + 10°	37.3	27.7	37.5	34.2
20m + 0°	31.9	23.7	36.8	30.8
20m + 5°	32.2	23.3	36.5	30.7
20m + 10°	33.2	23.3	36.6	31.0

TABLE 4.4: MapQR+HRMapNet performance under large-scale initial pose errors. Results show individual category performance and overall mAP under substantial localization uncertainties.

Translation / Rotation	**$AP_{div.}$**	**$AP_{ped.}$**	**$AP_{bou.}$**	**mAP**
0m + 0°	73.5	72.1	72.7	72.7
0m + 5°	66.8	68.8	67.4	67.7
0m + 10°	61.1	64.5	60.7	62.1
5m + 0°	50.5	47.3	49.7	49.2
5m + 5°	49.6	47.6	49.3	48.8
5m + 10°	49.9	45.8	48.0	47.9
10m + 0°	44.8	37.9	44.1	42.3
10m + 5°	44.8	38.1	44.9	42.6
10m + 10°	45.9	37.7	44.9	42.8
20m + 0°	43.4	36.9	45.5	41.9
20m + 5°	43.7	36.6	46.1	42.1
20m + 10°	44.3	37.2	46.7	42.7

Figure 4.7(c) provides a comprehensive bar chart comparison of both methods across twelve different large-scale perturbation conditions, revealing critical insights about HD map-

ping system behavior under significant initial pose errors that severely impact both historical map maintenance and prior knowledge utilization.

The bar chart demonstrates that both systems experience sharp performance drops when translation errors reach 5 meters, representing a critical threshold in the effectiveness of prior map integration. According to Tables 4.3 and 4.4, MapTRv2+HRMapNet degrades from 67.1 mAP to 44.1 mAP and MapQR+HRMapNet drops from 72.7 mAP to 49.2 mAP, representing approximately 34% and 32% performance reduction respectively. This 5-meter threshold represents a critical breakdown point where the spatial misalignment between current observations and historical map data becomes too severe for effective integration, fundamentally compromising the core value proposition of prior map utilization in HD mapping systems.

At extreme perturbation levels (20m + 10°), the performance gap between methods becomes even more pronounced. As detailed in the tables, MapQR+HRMapNet maintains 42.7 mAP while MapTRv2+HRMapNet degrades to 31.0 mAP, representing a substantial 11.7-point advantage. This demonstrates that even when initial pose errors catastrophically compromise both map updating accuracy and prior map retrieval effectiveness, MapQR's architecture maintains significantly better HD mapping capabilities, suggesting superior learned representations that can partially compensate for spatial misalignment. Although map construction is essentially a point set prediction task, MapQR utilizes instance queries rather than point queries. These instance queries are scattered for the prediction of point sets and subsequently gathered for the final matching. The base map instance queries are scattered to different reference points and added with positional embeddings, then these scattered queries are gathered back to enhance information within each map instance [27].

The large-scale results also show that rotation errors exhibit diminishing impact at large translation scales. Both tables confirm that when translation errors exceed 10 meters, the additional impact of rotation perturbations becomes minimal, indicating that spatial displacement errors dominate over orientation errors in disrupting the historical map integration process.

Semantic Category-Specific Robustness Analysis

The detailed performance breakdown across map element categories in Tables 4.3 and 4.4 reveals critical insights into how different HD map semantic elements respond to pose perturbations that affect prior map utilization effectiveness.

Pedestrian(ped.) crossing detection exhibits the most severe degradation under pose perturbations across both baseline methods. Table 4.3 shows pedestrian performance degrading from 65.4 mAP at perfect localization to 23.3 mAP under extreme conditions (20m + 10°), representing a catastrophic 64% performance loss. Similarly, Table 4.4 demonstrates pedestrian prediction dropping from 72.1 mAP to 37.2 mAP, indicating a 48% degradation. This heightened vulnerability reflects the discrete, spatially-precise nature of pedestrian crossings, which require exact spatial correspondence between current observations and historical map data for successful detection. When pose errors disrupt this correspondence, the prior map information becomes counterproductive rather than beneficial for crosswalk detection.

Road boundary(bou.) detection demonstrates superior robustness to pose perturbations

across both architectures. Table 4.3 shows boundary detection maintaining relatively stable performance from 68.4 mAP to 36.6 mAP, representing a 46% degradation, while Table 4.4 preserves boundary performance from 72.7 mAP to 46.7 mAP, showing a 36% degradation. This resilience stems from the continuous, extended spatial characteristics of road boundaries, which provide redundant spatial cues that maintain partial utility even when pose errors misalign historical map information. The geometric continuity of boundaries enables the HD mapping system to leverage prior map knowledge effectively even under significant spatial displacement.

Lane divider(div.) detection exhibits intermediate robustness characteristics, with performance patterns falling between the extreme vulnerability of pedestrian crossings and the resilience of boundaries. Table 4.3 shows divider performance decreasing from 67.5 mAP to 33.2 mAP, representing a 51% degradation, while Table 4.4 demonstrates degradation from 73.5 mAP to 44.3 mAP, showing a 40% degradation. This intermediate sensitivity reflects the linear but segmented nature of lane dividers, which provides some spatial redundancy for prior map utilization but lacks the extensive coverage that makes boundary detection robust to spatial misalignment.

The comparative analysis also reveals that MapQR+HRMapNet's architectural advantages are most pronounced for the most vulnerable semantic category. At extreme perturbation levels (20m + 10°), MapQR+HRMapNet maintains 37.2 mAP for Ped compared to MapTRv2+HRMapNet's 23.3 mAP, representing a 60% performance advantage precisely where prior map integration faces the greatest challenges. This demonstrates that MapQR's [56] query-based attention mechanisms provide superior capability for maintaining semantic detection performance when historical map information becomes spatially misaligned, offering critical advantages for safety-critical HD mapping applications where localization quality cannot be guaranteed.

Visualization Analysis

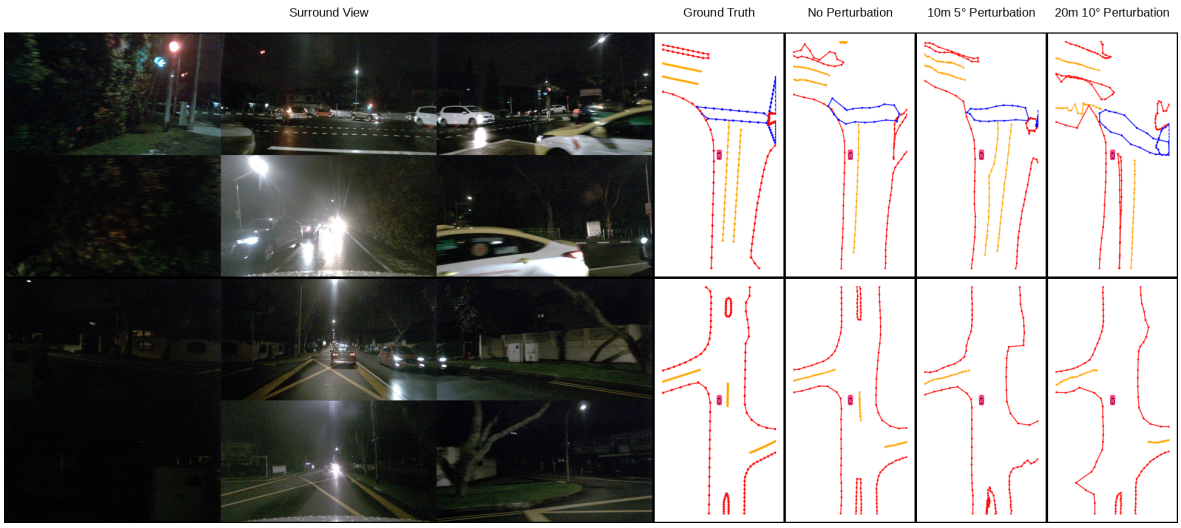

FIGURE 4.8: MapTRv2+HRMapNet visualization results on Nuscenes dataset with different test perturbation of night scenario

Figures 4.8, 4.9, and 4.10 present qualitative comparisons across three challenging environmental scenarios where historical map integration provides substantial benefits under ideal localization conditions but demonstrates vulnerability when subjected to pose perturbations.

The visualization analysis across night scenarios in Figure 4.8 demonstrates the fundamental challenge that poor lighting conditions pose for online map perception systems. Under perfect localization conditions (No Perturbation), the integration of historical rasterized maps enables the system to maintain robust detection of road boundaries, lane dividers, and pedestrian crossings despite severe visibility limitations that would otherwise compromise sensor-only perception. The surround view images clearly show challenging lighting conditions with limited visibility of lane markings and road structures, yet the ground truth comparison reveals that historical map integration successfully compensates for these sensory limitations.

However, the progressive degradation becomes evident as pose perturbations increase. Under 10m 5°perturbation conditions, the visualization reveals noticeable misalignment between predicted map elements and their true spatial positions, with particular deterioration in the detection of discrete elements such as pedestrian crossings. The 20m 10°perturbation results demonstrate severe spatial misalignment where the historical map information becomes counterproductive, leading to false detections and missed critical road infrastructure. This degradation occurs not merely due to noise interference but fundamentally because the spatial displacement causes the original 200m×100m historical map segment to no longer adequately cover the current perception region, resulting in incomplete prior knowledge for map reconstruction.

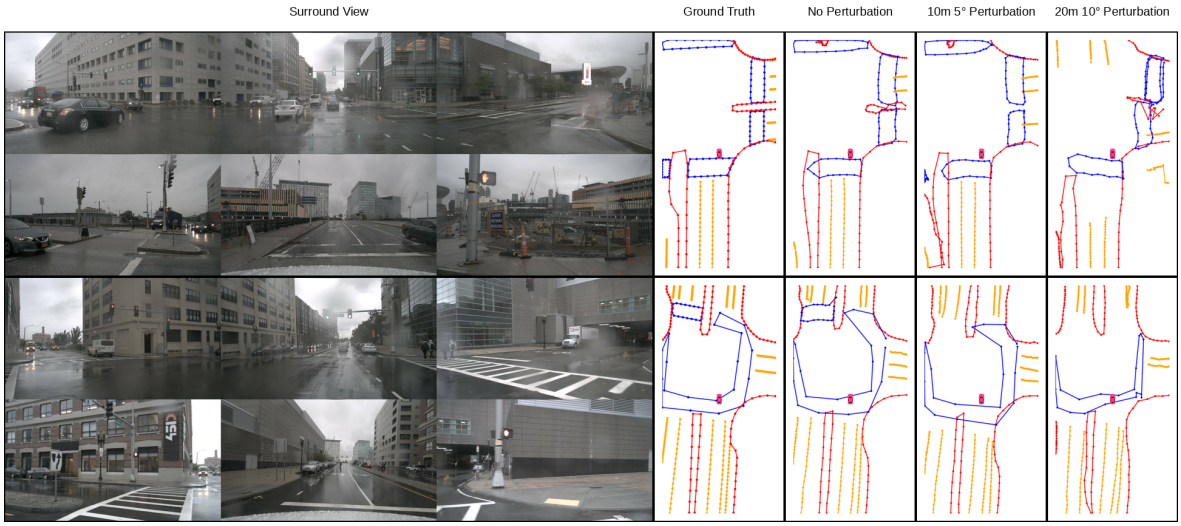

FIGURE 4.9: MapTRv2+HRMapNet visualization results on Nuscenes dataset with different test perturbation of rainy scenario

The rainy scenario analysis in Figure 4.9 reveals similar patterns of performance degradation under adverse weather conditions. The surround view images demonstrate how precipitation affects visual perception quality, creating additional challenges for feature detection and spatial reasoning. Under ideal pose conditions, the historical map integration maintains high-quality vectorized map generation despite the compromised visual input. The ground truth comparison shows accurate detection of complex road geometry including curved boundaries

and intersection configurations that would be difficult to perceive reliably from sensor data alone under these weather conditions.

The perturbation analysis reveals that rainy conditions compound the challenges of pose uncertainty, creating a multiplicative effect on mapping quality degradation. The 10m 5°perturbation results show more severe performance degradation compared to clear weather conditions, suggesting that environmental challenges reduce the system's tolerance for localization errors. At extreme perturbation levels (20m 10°), the combination of adverse weather and pose uncertainty leads to substantial mapping failures, with missed road boundaries and incorrectly positioned lane markings. This demonstrates that the overlapping semantic information between misaligned historical map segments becomes insufficient to support reliable map reconstruction when environmental conditions already stress the perception system.

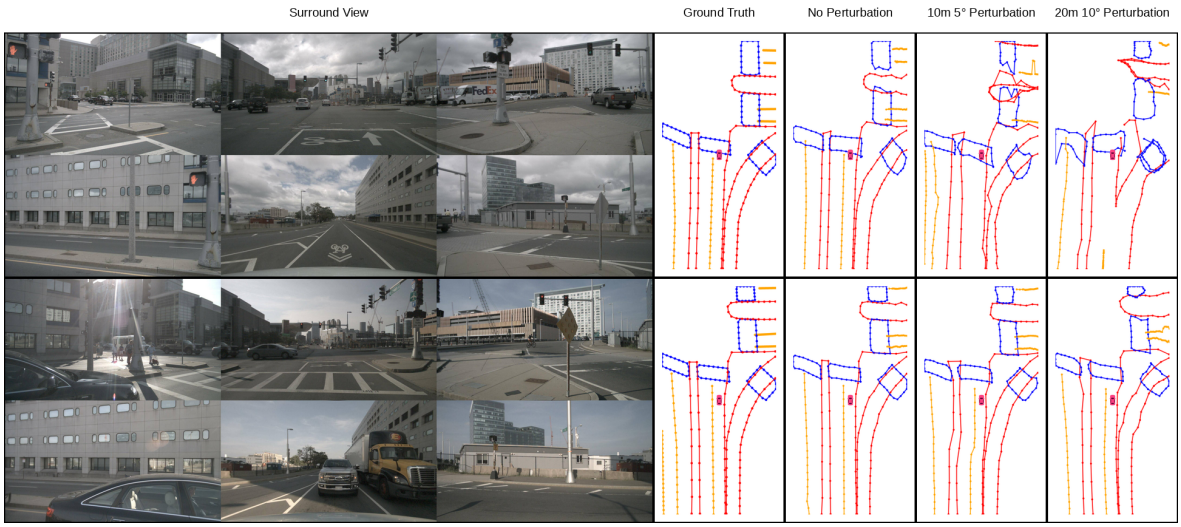

FIGURE 4.10: MapTRv2+HRMapNet visualization results on Nuscenes dataset with different test perturbation of occulusion scenario

The severe occlusion scenario in Figure 4.10 presents perhaps the most challenging conditions for online map perception, where large vehicles and infrastructure elements block critical road features from sensor observation. The surround view images show significant occlusion of road markings, boundaries, and pedestrian infrastructure that would render sensor-only mapping systems ineffective. Under perfect localization, the historical map integration demonstrates exceptional capability in maintaining complete road structure detection despite extensive sensory occlusion, validating the core value proposition of prior map utilization in HD mapping systems.

The visualization reveals that occlusion scenarios exhibit the most severe sensitivity to pose perturbations among the three challenging conditions examined. The 10m 5°perturbation already shows substantial degradation in mapping quality, with misaligned road boundaries and incomplete lane structure detection. This heightened sensitivity occurs because occlusion scenarios provide limited redundant visual information that could compensate for spatial misalignment between current observations and historical map data. Under extreme perturbation conditions (20m 10°), the mapping system fails to maintain coherent road structure representation, as the combination of sensory occlusion and spatial misalignment creates a condition where neither current sensor data nor historical map information

provides sufficient spatial context for reliable perception.

The comprehensive visualization analysis across these challenging scenarios reveals that pose perturbations could cause degradation in historical map integration. The primary effect involves direct spatial misalignment between current sensor observations and retrieved historical map segments, disrupting the correspondence necessary for effective map updating and retrieval. The secondary effect involves the loss of spatial coverage, where pose errors cause the historical 200m×100m map segments to inadequately cover the current perception region, leading to incomplete prior knowledge that cannot support comprehensive map reconstruction.

Chapter 5

Integration Framework Design

The traditional paradigm of HD mapping and localization processing fundamentally limits system performance through error propagation and suboptimal resource utilization. Contemporary research demonstrates that the intrinsic coupling between spatial perception and pose estimation creates opportunities for joint optimization that significantly exceed the capabilities of independently designed components. This architectural evolution represents more than incremental improvement; it constitutes a fundamental reconceptualization of how autonomous systems interact with their spatial environment.

The mathematical foundation for this integration lies in the shared optimization landscape where mapping accuracy and localization precision are inherently interdependent. When vehicle pose uncertainty propagates into map construction errors, these errors subsequently degrade future localization attempts, creating a compounding effect that limits long-term system performance. Recent work such as RTMap [12] demonstrates that unified frameworks can address this challenge by treating pose estimation and map construction as coupled optimization problems. As shown in their experimental results, such joint approaches can achieve improved performance compared to sequential processing, particularly when dealing with map updates and change detection scenarios. Building on these insights, our proposed framework explores similar integration strategies for HD mapping and localization tasks.

5.1 End-to-End Online Processing Architecture

5.1.1 Multi-View Camera Input and BEV Encoder

The architecture employs a surround-view camera configuration following established practices from HDMapNet [19] and BEVFormer [22]. This unified input strategy captures comprehensive 360-degree environmental coverage while maintaining computational efficiency through parallel feature extraction across all viewpoints, providing the semantic richness necessary for both HD mapping and precise localization tasks.

The BEV encoder backbone transforms multi-perspective camera observations into unified bird's-eye view representations, addressing the fundamental challenge of viewpoint dependency in multi-camera systems. The transformer-based view transformation projects

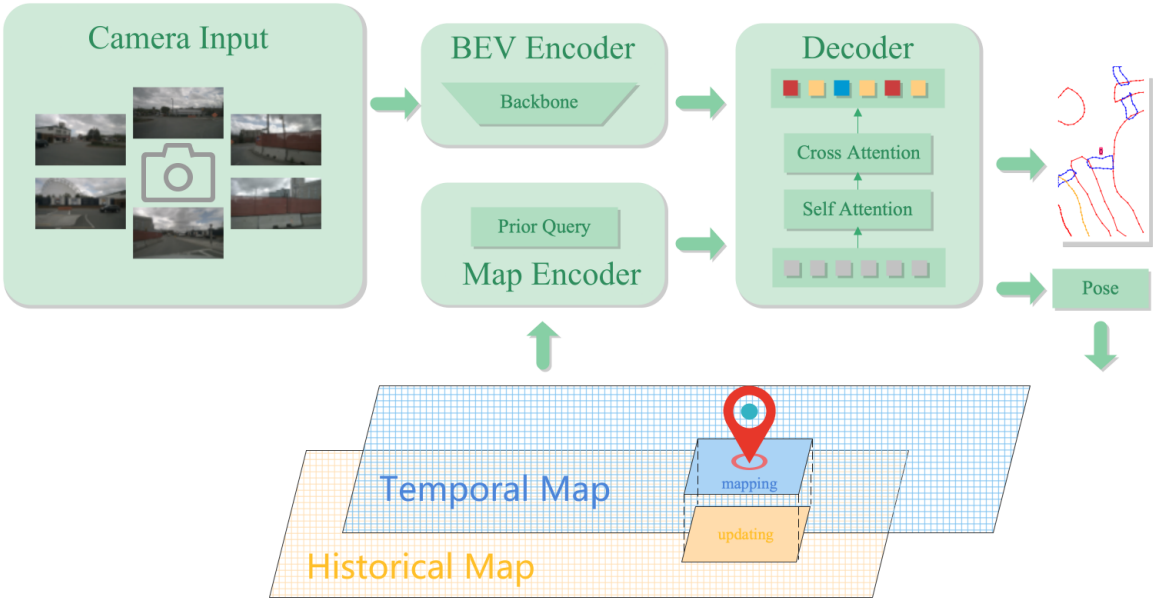

FIGURE 5.1: End-to-end architecture for simultaneous HD mapping and localization. Multi-view camera inputs are processed through a BEV encoder backbone while hybrid prior maps undergo representation-specific encoding. The novel two-stage coarse-to-fine decoder leverages rasterized temporal maps for rapid initial estimation and vectorized historical maps for precision refinement, generating dual outputs: vectorized HD maps and 6-DOF pose estimates. The offline dual-layer storage system maintains vectorized historical maps and rasterized temporal maps with intelligent tile-indexed organization.

perspective features into metrically consistent overhead coordinates, eliminating perspective distortions that would otherwise accumulate during map fusion operations, particularly for distant objects where geometric accuracy proves critical for precise localization.

5.1.2 Hybrid Map Representation Integration

Building upon the rasterization philosophy demonstrated in MapVR [55], which reveals that "integrating the philosophy of rasterization into map vectorization" significantly enhances performance through superior sensitivity to geometric deviations, we propose a novel hybrid representation strategy that leverages the complementary advantages of both vectorized and rasterized map formats across different temporal scales.

Representation Selection Rationale: Contemporary localization approaches face a fundamental trade-off between geometric precision and neural network compatibility. While BEV-Locator [57] achieves superior lateral pose estimation using vectorized HD maps, it operates as an unexplainable black-box with limited yaw angle estimation performance. Conversely, rasterized representations enable explainable dense correspondence establishment but sacrifice geometric precision through discretization.

Our hybrid approach addresses these limitations through intelligent representation selection based on temporal characteristics and processing requirements. Following MapVR's insights that rasterization provides "precise and geometry-aware supervision" while main-

taining computational efficiency, we design a representation-aware processing pipeline that optimizes both accuracy and explainability.

Temporal-Aware Map Encoding: The map encoder processes different representations optimized for their respective temporal and geometric characteristics:

$$M_{encoded} = \begin{cases} \text{VectorEncoder}(M_{historical}) & \text{for stable geometric precision} \\ \text{RasterEncoder}(M_{temporal}) & \text{for dynamic dense correspondence} \end{cases} \tag{5.1}$$

This selective encoding strategy enables optimal representation utilization while maintaining computational efficiency through specialized processing pathways tailored to each format's strengths.

5.1.3 Novel Representation-Aware Coarse-to-Fine Decoder

Drawing inspiration from MapLocNet's coarse-to-fine registration strategy [48] and MapQR's query-based attention mechanisms [56], we propose a novel two-stage architecture that naturally leverages our hybrid representation storage system to achieve progressive accuracy refinement. Unlike traditional approaches that utilize different network feature scales, our method exploits the inherent complementary characteristics of temporal and historical map representations.

Architectural Philosophy: The design addresses the fundamental trade-off between responsiveness and precision by utilizing representation-specific strengths: rasterized temporal maps provide rapid dense correspondence for coarse estimation, while vectorized historical maps enable geometric precision for fine refinement. This approach eliminates the need for multi-scale feature storage while enabling focused computation on relevant spatial regions.

Stage 1 - Temporal-Based Coarse Estimation: The coarse stage utilizes rasterized temporal maps to establish rapid initial spatial correspondences:

$$[M_{coarse}, P_{coarse}] = \text{CoarseDecoder}(F_{BEV}, M_{temporal}) \tag{5.2}$$

where $M_{temporal}$ represents recently observed rasterized map information that enables efficient dense correspondence matching through grid-based attention mechanisms. The rasterized format facilitates comprehensive spatial coverage while maintaining computational efficiency, providing robust initialization that effectively handles dynamic environmental conditions.

The temporal-based approach leverages pixel-wise correspondences to establish broad spatial relationships between current BEV observations and recent environmental changes. This stage prioritizes coverage and responsiveness over precision, generating spatial attention priors that guide subsequent fine-stage processing toward relevant geometric regions.

Stage 2 - Historical-Based Fine Refinement: The fine stage incorporates vectorized

historical maps through pose-guided rasterization to achieve sub-pixel geometric precision:

$$M_{refined} = \text{PoseGuidedRasterize}(M_{historical}, P_{coarse}, \sigma_{fine}) \tag{5.3}$$

$$[M_{fine}, P_{fine}] = \text{FineDecoder}(F_{BEV}, M_{refined}, P_{coarse}) \tag{5.4}$$

where P_{coarse} provides spatial attention guidance and σ_{fine} determines the resolution requirements for precision refinement. The pose-guided rasterization utilizes MapVR's differentiable rasterization approach [55] to convert relevant portions of vectorized historical maps while preserving geometric precision essential for accurate localization.

5.2 Offline Map Maintenance Architecture

5.2.1 Dual-Layer Hybrid Representation Storage

Following the dual-layer architectural concept from Uni-PrevPredMap [30], we propose a novel hybrid representation storage system that strategically leverages different map formats to optimize the trade-off between storage efficiency and neural processing compatibility across temporal scales.

Historical Layer - Vectorized Storage: Long-term stable infrastructure elements are maintained in vectorized format to preserve geometric precision and storage efficiency. Vectorized elements support complex geometric operations, semantic queries, and precise spatial relationships while enabling conservative update policies that preserve map integrity across extended temporal periods.

Temporal Layer - Rasterized Storage: Recent observations and dynamic environmental changes are stored in rasterized format to enable direct neural network processing without conversion overhead. This representation facilitates rapid dense correspondence matching, real-time map updates, and change detection operations while maintaining native compatibility with transformer-based architectures.

This hybrid storage strategy exploits distinct temporal-accuracy characteristics: vectorized historical maps provide persistent geometric constraints with storage efficiency, while rasterized temporal maps ensure neural processing efficiency with immediate availability. This separation enables optimal resource utilization while preserving both long-term stability and short-term responsiveness.

5.2.2 Hybrid Tile-Indexed Storage and Retrieval Method

Extending the tile-indexed approach from Uni-PrevPredMap [30], our offline system employs sophisticated spatial indexing that accommodates both vectorized and rasterized

map formats within unified tile structures. Each tile represents a discrete geographical region indexed through Universal Transverse Mercator (UTM) coordinates $(i, j) = \lfloor (UTM_{east}, UTM_{north})/l \rfloor$, where l denotes the tile dimension corresponding to the perception range.

Unified Tile Organization: Each tile maintains both representation layers within consistent spatial structures:

$$\text{Tile}(i, j) = \{\mathcal{H}_{vector}(i, j), \mathcal{T}_{raster}(i, j)\} \tag{5.5}$$

The historical vectorized layer preserves lane centerlines, boundary polygons, and traffic sign locations as continuous mathematical objects with parametric curves and precise coordinate references. The temporal rasterized layer stores recent observations as dense grids aligned with identical UTM coordinate systems, maintaining spatial consistency with vectorized elements while enabling direct neural processing.

Spatial Coverage and Retrieval: The retrieval mechanism employs UTM coordinate analysis to ensure complete spatial coverage for both representation layers through boundary-aware adjacency selection:

$$I = \begin{cases} \{i_t - 1, i_t\} & \text{if } UTM_{east} \bmod l < l/2 \\ \{i_t\} & \text{if } UTM_{east} \bmod l = l/2 \\ \{i_t, i_t + 1\} & \text{if } UTM_{east} \bmod l > l/2 \end{cases} \tag{5.6}$$

$$J = \begin{cases} \{j_t - 1, j_t\} & \text{if } UTM_{north} \bmod l < l/2 \\ \{j_t\} & \text{if } UTM_{north} \bmod l = l/2 \\ \{j_t, j_t + 1\} & \text{if } UTM_{north} \bmod l > l/2 \end{cases} \tag{5.7}$$

This selection strategy ensures comprehensive environmental information retrieval for both vectorized historical data and rasterized temporal observations, preventing perception gaps while maintaining computational efficiency through spatial locality principles.

Chapter 6

Discussion

6.1 Research Contributions and Impact

This thesis makes several significant contributions to the autonomous driving research community. First, it provides the first comprehensive survey that systematically bridges HD mapping and localization domains, establishing a unified taxonomic framework for comparative analysis. The systematic categorization of methods based on prior information utilization and architectural approaches provides a structured foundation for future research directions. Second, the identification of optimal integration points where mapping and localization tasks mutually reinforce rather than constrain each other establishes critical design principles for next-generation autonomous systems. The demonstration that both domains converge on similar technical foundations validates the feasibility of unified architectural approaches while highlighting the importance of semantic-geometric correspondence learning. Third, the experimental analysis provides quantitative evidence for the critical importance of realistic uncertainty modeling in both training and evaluation protocols. The establishment of performance degradation thresholds and robustness characteristics across different environmental conditions provides practical guidance for system deployment and operational safety requirements.

6.2 Limitations and Future Work

Several limitations of current approaches emerged through this analysis. The conservative perturbation assumptions employed in existing methods, typically limited to 2-5 meter uncertainties, inadequately simulate realistic GNSS deviations that frequently exceed 10-20 meters in challenging urban environments. This discrepancy between experimental validation conditions and actual deployment scenarios suggests that while current methods demonstrate excellent refinement capabilities, their robustness for global relocalization scenarios remains inadequately characterized. The computational complexity scaling of multi-modal approaches presents challenges for real-time deployment, particularly for cost-sensitive applications where sophisticated sensor synchronization and processing pipelines may limit practical adoption. The trade-off between localization accuracy and system complexity requires continued inves-

tigation to achieve optimal performance-efficiency balance. Future research should prioritize the development of adaptive integration strategies that can dynamically adjust processing based on available prior information and environmental conditions. The investigation of collaborative mapping systems where vehicles contribute to and benefit from shared historical knowledge represents a promising direction for scalable autonomous driving deployment. Additionally, the exploration of learned prior representations and hybrid approaches combining multiple types of prior information could yield further performance improvements.

6.3 Conclusion

The convergence of learning-based HD mapping and localization technologies represents a fundamental evolution in autonomous driving system architecture. This thesis demonstrates that the traditional separation between mapping and localization processes constitutes an artificial constraint that limits system performance and adaptability. The systematic analysis reveals that integrated approaches can achieve superior accuracy, robustness, and efficiency compared to sequential processing paradigms. The architectural insights and experimental findings presented in this work provide a foundation for developing next-generation autonomous driving systems capable of maintaining precision while adapting to dynamic operational environments. The unified framework for understanding mapping-localization integration will inform future research and development efforts toward truly end-to-end autonomous driving capabilities that can operate reliably across diverse and challenging real-world conditions. The successful completion of this thesis, conducted through collaborative efforts between academic research at the University of Twente and industry expertise at Scania, demonstrates the value of bridging theoretical advancement with practical deployment requirements. The findings contribute to the broader goal of developing scalable, safe, and efficient autonomous transportation technologies that can adapt to the evolving demands of modern mobility systems.

Bibliography

[1] Kaleab Taye Asrat and Hyung-Ju Cho. "A Comprehensive Survey on High-Definition Map Generation and Maintenance". In: *ISPRS International Journal of Geo-Information* 13.7 (2024), p. 232 (cit. on pp. 2, 3).

[2] Josep Aulinas, Yvan Petillot, Joaquim Salvi, and Xavier Lladó. "The SLAM problem: a survey". In: *Artificial Intelligence Research and Development* (2008), pp. 363–371 (cit. on p. 2).

[3] Zhibin Bao, Sabir Hossain, Haoxiang Lang, and Xianke Lin. "High-definition map generation technologies for autonomous driving". In: *arXiv preprint arXiv:2206.05400* (2022) (cit. on p. 9).

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. "nuScenes: A multimodal dataset for autonomous driving". In: *CVPR 2020*. 2020 (cit. on p. 10).

[5] Andrea Boscolo Camiletto, Alfredo Bochicchio, Alexander Liniger, Dengxin Dai, and Abel Gawel. "U-bev: Height-aware bird's-eye-view segmentation and neural map-based relocalization". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 5597–5604 (cit. on pp. 18, 19).

[6] Athanasios Chalvatzaras, Ioannis Pratikakis, and Angelos A. Amanatiadis. "A Survey on Map-Based Localization Techniques for Autonomous Vehicles". In: *IEEE Transactions on Intelligent Vehicles* 8.2 (2023), pp. 1574–1596. DOI: 10.1109/TIV.2022.3192102 (cit. on pp. 1, 2, 34).

[7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva 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 (cit. on p. 10).

[8] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. "Deep learning for visual localization and mapping: A survey". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023) (cit. on p. 7).

[9] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. "Maptracker: Tracking with strided memory fusion for consistent vector hd mapping". In: *European Conference on Computer Vision*. Springer. 2024, pp. 90–107 (cit. on pp. 24, 25).

[10] Sehwan Choi, Jungho Kim, Hongjae Shin, and Jun Won Choi. "Mask2map: Vectorized hd map construction using bird's eye view segmentation masks". In: *European Conference on Computer Vision*. Springer. 2024, pp. 19–36 (cit. on pp. 22, 23).

[11] Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. "Pivotnet: Vectorized pivot learning for end-to-end hd map construction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3672–3682 (cit. on pp. 22, 23).

[12] Yuheng Du, Sheng Yang, Lingxuan Wang, Zhenghua Hou, Chengying Cai, Zhitao Tan, Mingxia Chen, Shi-Sheng Huang, and Qiang Li. "RTMap: Real-Time Recursive Mapping with Change Detection and Localization". In: *arXiv preprint arXiv:2507.00980* (2025) (cit. on pp. 29, 51).

[13] Gamal Elghazaly, Raphaël Frank, Scott Harvey, and Stefan Safko. "High-definition maps: Comprehensive survey, challenges, and future perspectives". In: *IEEE Open Journal of Intelligent Transportation Systems* 4 (2023), pp. 527–550 (cit. on pp. 1–4, 7–9).

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. "Vision meets robotics: The kitti dataset". In: *The international journal of robotics research* 32.11 (2013), pp. 1231–1237 (cit. on p. 11).

[15] Yuzhe He, Shuang Liang, Xiaofei Rui, Chengying Cai, and Guowei Wan. "Egovm: Achieving precise ego-localization using lightweight vectorized maps". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 12248–12255 (cit. on pp. 14–17).

[16] Zhou Jiang, Zhenxin Zhu, Pengfei Li, Huan-ang Gao, Tianyuan Yuan, Yongliang Shi, Hang Zhao, and Hao Zhao. "P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors". In: *IEEE Robotics and Automation Letters* (2024) (cit. on pp. 26, 30).

[17] Nayeon Kim, Hongje Seong, Daehyun Ji, and Sujin Jang. "Unveiling the Hidden: Online Vectorized HD Map Construction with Clip-Level Token Interaction and Propagation". In: *arXiv preprint arXiv:2411.11002* (2024) (cit. on p. 25).

[18] John Lambert and James Hays. "Trust, but Verify: Cross-Modality Fusion for HD Map Change Detection". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (cit. on pp. 9, 10).

[19] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. "Hdmapnet: A local semantic map learning and evaluation framework". In: *arXiv preprint arXiv:2107.06307* 1.3 (2021), p. 5 (cit. on pp. 3, 21–23, 30, 51).

[20] Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. "Lanesegnet: Map learning with lane segment perception for autonomous driving". In: *International Conference on Learning Representations (ICLR)* (2024) (cit. on p. 23).

[21] Xiaofan Li, Zhihao Xu, Chenming Wu, Zhao Yang, Yumeng Zhang, Jiang-Jiang Liu, Haibao Yu, Fan Duan, Xiaoqing Ye, Yuan Wang, et al. "U-ViLAR: Uncertainty-Aware Visual Localization for Autonomous Driving via Differentiable Association and Registration". In: *arXiv preprint arXiv:2507.04503* (2025) (cit. on pp. 15–17).

[22] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024) (cit. on p. 51).

[23] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. "Maptr: Structured modeling and learning for online vectorized hd map construction". In: *International Conference on Learning Representations (ICLR)* (2022) (cit. on pp. 3, 22, 23).

[24] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. "Maptrv2: An end-to-end framework for online vectorized hd map construction". In: *International Journal of Computer Vision* (2024), pp. 1–23 (cit. on pp. 22, 23, 30, 43).

[25] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. "Mgmap: Mask-guided learning for online vectorized hd map construction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14812–14821 (cit. on pp. 22, 23).

[26] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. "Vectormapnet: End-to-end vectorized hd map learning". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 22352–22369 (cit. on pp. 3, 30).

[27] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. "Leveraging enhanced queries of point sets for vectorized map construction". In: *European Conference on Computer Vision*. Springer. 2024, pp. 461–477 (cit. on pp. 22, 23, 43, 46).

[28] Jinyu Miao, Tuopu Wen, Ziang Luo, Kangan Qian, Zheng Fu, Yunlong Wang, Kun Jiang, Mengmeng Yang, Jin Huang, Zhihua Zhong, et al. "Efficient End-to-end Visual Localization for Autonomous Driving with Decoupled BEV Neural Matching". In: *arXiv preprint arXiv:2503.00862* (2025) (cit. on pp. 6, 14–17).

[29] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. "ORB-SLAM: A versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163 (cit. on p. 3).

[30] Nan Peng, Xun Zhou, Mingming Wang, Guisong Chen, and Wenqi Xu. "UniPrevPredMap: Extending PrevPredMap to a Unified Framework of Prior-Informed Modeling for Online Vectorized HD Map Construction". In: *arXiv preprint arXiv:2504.06647* (2025) (cit. on pp. 25, 28, 29, 31, 54).

[31] Nan Peng, Xun Zhou, Mingming Wang, Xiaojun Yang, Songming Chen, and Guisong Chen. "Prevpredmap: Exploring temporal modeling with previous predictions for online vectorized hd map construction". In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2025, pp. 8134–8143 (cit. on pp. 3, 24, 25).

[32] Iván Puente, H González-Jorge, J Martínez-Sánchez, and Pedro Arias. "Review of mobile mapping and surveying technologies". In: *Measurement* 46.7 (2013), pp. 2127–2145 (cit. on pp. 4, 7, 9).

[33] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. "End-to-end vectorized hd-map construction with piecewise bezier curve". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 13218–13228 (cit. on pp. 22, 23).

[34] Tyler G Reid, Sarah E Houts, Robert Cammarata, Graham Mills, Siddharth Agarwal, Ankit Vora, and Gaurav Pandey. “Localization Requirements for Autonomous Vehicles”. In: *SAE International Journal of Connected and Automated Vehicles* 2.12-02-03-0012 (2019) (cit. on p. 1).

[35] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios 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. 21632–21642 (cit. on pp. 11, 17–19, 34, 35).

[36] Yujiao Shi and Hongdong Li. “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*. 2022, pp. 17010–17020 (cit. on p. 34).

[37] Juyeb Shin, Hyeonjun Jeong, Francois Rameau, and Dongsuk Kum. “Instagram: Instance-level graph modeling for vectorized hd map learning”. In: *IEEE Transactions on Intelligent Transportation Systems* (2025) (cit. on pp. 21, 23).

[38] Jürgen Sturm, Wolfram Burgard, and Daniel Cremers. “Evaluating egomotion and structure-from-motion approaches using the TUM RGB-D benchmark”. In: *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*. Vol. 13. 2012, p. 6 (cit. on p. 11).

[39] Rémy Sun, Li Yang, Diane Lingrand, and Frédéric Precioso. “Mind the map! Accounting for existing map information when estimating online HDMaps from sensor”. In: *arXiv preprint arXiv:2311.10517* (2023) (cit. on pp. 9, 27).

[40] Hamid Taheri and Zhao Chun Xia. “SLAM; definition and evolution”. In: *Engineering Applications of Artificial Intelligence* 97 (2021), p. 104032 (cit. on pp. 2, 4, 8).

[41] Rongxuan Wang, Xin Lu, Xiaoyang Liu, Xiaoyi Zou, Tongyi Cao, and Ying Li. “Priormapnet: Enhancing online vectorized hd map construction with priors”. In: *arXiv preprint arXiv:2408.08802* (2024) (cit. on pp. 24, 25).

[42] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao. “Stream Query Denoising for Vectorized HD-Map Construction”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 203–220 (cit. on pp. 24, 25).

[43] Benny Wijaya, Kun Jiang, Mengmeng Yang, Tuopu Wen, Yunlong Wang, Xuewei Tang, Zheng Fu, Gracelynn Soesanto, Taohua Zhou, Jinyu Miao, et al. “High Definition Map Mapping and Update: A General Overview and Future Directions”. In: () (cit. on p. 12).

[44] Benny Wijaya, Kun Jiang, Mengmeng Yang, Tuopu Wen, Yunlong Wang, Xuewei Tang, Zheng Fu, Taohua Zhou, and Diange Yang. “High Definition Map Mapping and Update: A General Overview and Future Directions”. In: *arXiv preprint arXiv:2409.09726* (2024) (cit. on pp. 2, 8).

[45] Lena Wild, Ludvig Ericson, Rafael Valencia, and Patric Jensfelt. “Exelmap: Explainable element-based hd-map change detection and update”. In: *ECCV 2024 2nd VCAD Workshop* (2024) (cit. on p. 9).

[46] Lena Wild, Rafael Valencia, and Patric Jensfelt. “ArgoTweak: Towards Self-Updating HD Maps through Structured Priors”. In: *ICCV 2025* (2025) (cit. on p. 11).

[47] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. “Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2023 (cit. on p. 10).

[48] Hang Wu, Zhenghao Zhang, Siyuan Lin, Xiangru Mu, Qiang Zhao, Ming Yang, and Tong Qin. “Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 13198–13205 (cit. on pp. 17–19, 53).

[49] Xin Xia, Zonglin Meng, Xu Han, Hanzhao Li, Takahiro Tsukiji, Runsheng Xu, Zhaoliang Zheng, and Jiaqi Ma. “An automated driving systems data acquisition and analytics platform”. In: *Transportation research part C: emerging technologies* 151 (2023), p. 104120 (cit. on p. 9).

[50] Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. “Neural map prior for autonomous driving”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 17535–17544 (cit. on pp. 3, 26, 30, 31).

[51] Zhenhua Xu, Kwan-Yee K. Wong, and Hengshuang Zhao. “InsMapper: Exploring inner-instance information for vectorized HD mapping”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 296–312 (cit. on pp. 22, 23).

[52] Jing Yang, Minyue Jiang, Sen Yang, Xiao Tan, Yingying Li, Errui Ding, Hanli Wang, and Jingdong Wang. “MGMapNet: Multi-Granularity Representation Learning for End-to-End Vectorized HD Map Construction”. In: *arXiv preprint arXiv:2410.07733* (2024) (cit. on pp. 22, 23).

[53] Jing Yang, Sen Yang, Xiao Tan, and Hanli Wang. “Histrackmap: Global vectorized high-definition map construction via history map tracking”. In: *arXiv preprint arXiv:2503.07168* (2025) (cit. on p. 25).

[54] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. “Streammapnet: Streaming mapping network for vectorized online hd map construction”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7356–7365 (cit. on pp. 3, 24, 25, 30).

[55] Gongjie Zhang, Jiahao Lin, Shuang Wu, Yilin Song, Zhipeng Luo, Yang Xue, Shijian Lu, and Zuoguan Wang. “Online Map Vectorization for Autonomous Driving: A Rasterization Perspective”. In: *arXiv preprint arXiv:2306.10502* (2023) (cit. on pp. 22, 23, 52, 54).

[56] Xiaoyu Zhang, Guangwei Liu, Zihao Liu, Ningyi Xu, Yunhui Liu, and Ji Zhao. “Enhancing vectorized map perception with historical rasterized maps”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 422–439 (cit. on pp. 3, 24, 26, 30, 31, 43, 47, 53).

[57] Zhihuang Zhang, Meng Xu, Wenqiang Zhou, Tao Peng, Liang Li, and Stefan Poslad. "Bev-locator: An end-to-end visual semantic localization network using multi-view images". In: *Science China Information Sciences* 68.2 (2025), p. 122106 (cit. on pp. 3, 6, 14–17, 52).

[58] Zhihuang Zhang, Jintao Zhao, Changyao Huang, and Liang Li. "Learning visual semantic map-matching for loosely multi-sensor fusion localization of autonomous vehicles". In: *IEEE Transactions on Intelligent Vehicles* 8.1 (2022), pp. 358–367 (cit. on p. 6).

[59] Junhui Zhao, Jingyue Shi, and Li Zhuo. "BEV perception for autonomous driving: State of the art and future perspectives". In: *Expert Systems with Applications* 258 (2024), p. 125103 (cit. on p. 3).

[60] Lili Zhao, Zhili Liu, Qian Yin, Lei Yang, and Meng Guo. "Towards Robust Visual Localization Using Multi-View Images and HD Vector Map". In: *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2024, pp. 814–820 (cit. on pp. 15–17).

[61] Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. "Himap: Hybrid representation learning for end-to-end vectorized hd map construction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 15396–15406 (cit. on pp. 22, 23).

[62] Zijie Zhou, Zhangshuo Qi, Luqi Cheng, and Guangming Xiong. "SegLocNet: Multimodal Localization Network for Autonomous Driving via Bird's-Eye-View Segmentation". In: *arXiv preprint arXiv:2502.20077* (2025) (cit. on pp. 3, 6, 17–19).

Appendix A

Appendix

In this appendix, we provide additional experimental results and materials to supplement the findings presented in the main text.

A.1 Localization Experiments Results

Localization experiments results from different train and test configuration matches.

Table A.1: Performance under ideal test conditions (Test: 0m, 0°)

Training Config	XY Pert. (m)	Yaw Pert. (°)	Lateral MAE(m)	Longitudinal MAE(m)	Yaw MAE(°)	Lateral R@1m(%)	Longitudinal R@1m(%)	Yaw R@1°(%)
kitti_0_0	0	0	0.352	0.322	2.513	94.43	95.20	33.68
kitti_10_5	10	5	0.672	1.067	2.468	76.53	48.50	31.48
kitti_20_0	20	0	0.389	0.344	2.272	92.31	94.21	34.50
kitti_20_10	20	10	0.677	1.066	2.396	76.53	48.65	32.05
kitti_30_10	30	10	0.701	1.082	2.391	73.91	47.52	32.56

Table A.2: Performance under moderate test perturbations (Test: 10m, 5°)

Training Config	XY Pert. (m)	Yaw Pert. (°)	Lateral MAE(m)	Longitudinal MAE(m)	Yaw MAE(°)	Lateral R@1m(%)	Longitudinal R@1m(%)	Yaw R@1°(%)
kitti_0_0	0	0	2.111	5.418	4.053	56.54	17.75	28.80
kitti_10_5	10	5	0.940	2.827	2.768	73.10	41.75	29.38
kitti_20_0	20	0	1.792	5.431	3.819	59.20	17.97	29.24
kitti_20_10	20	10	0.925	2.868	2.662	73.53	41.63	30.06
kitti_30_10	30	10	1.004	2.912	2.738	70.18	40.80	30.50

Table A.3: Performance under pure translational test perturbations (Test: 20m, 0°)

Training Config	XY Pert. (m)	Yaw Pert. (°)	Lateral MAE(m)	Longitudinal MAE(m)	Yaw MAE(°)	Lateral R@1m(%)	Longitudinal R@1m(%)	Yaw R@1°(%)
kitti_0_0	0	0	4.029	10.129	4.839	48.70	14.92	26.86
kitti_10_5	10	5	1.152	5.228	2.924	70.26	38.17	28.85
kitti_20_0	20	0	3.395	10.092	4.484	52.16	14.89	27.63
kitti_20_10	20	10	1.137	5.190	2.826	70.83	38.07	29.45
kitti_30_10	30	10	1.307	5.329	2.968	67.21	36.86	29.63

A.2 Mapping Experiments Visualization

More visualization results of complex environments in mapping experiments.

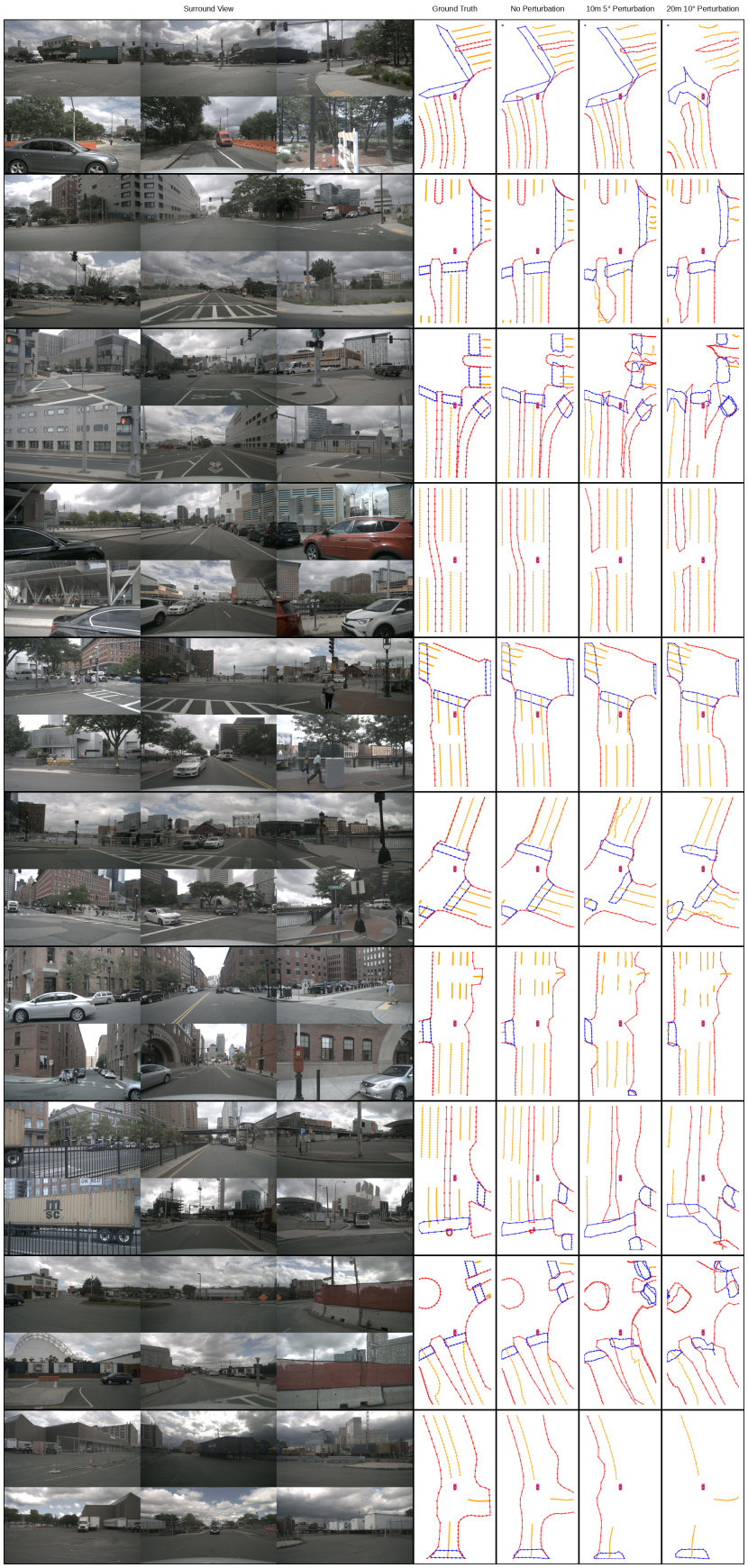

FIGURE A.1: Visualization results of mapping experiments

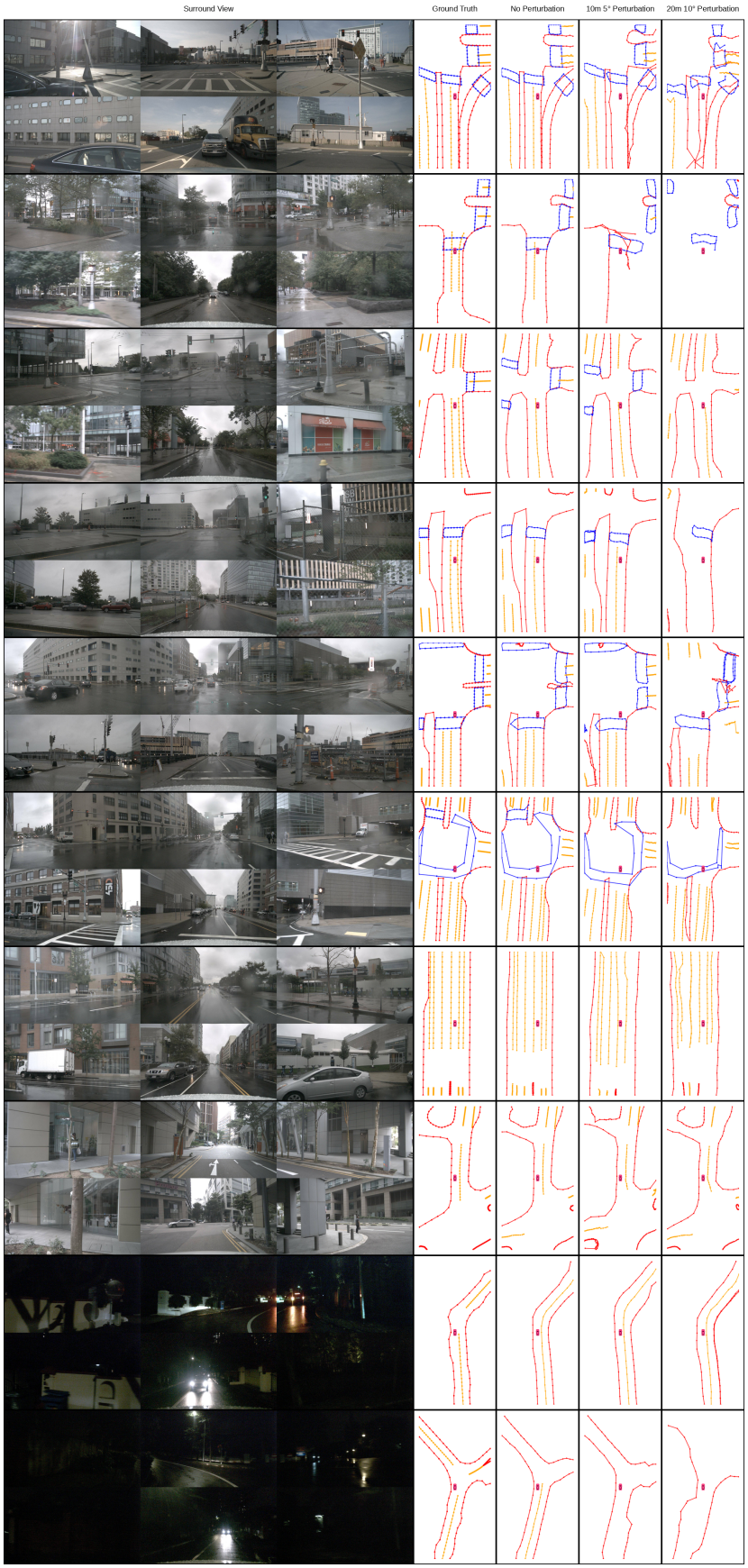

FIGURE A.2: Visualization results of mapping experiments

TABLE A.4: Performance under realistic test perturbations (Test: 20m, 10°)

Training Config	XY Pert. (m)	Yaw Pert. (°)	Lateral MAE(m)	Longitudinal MAE(m)	Yaw MAE(°)	Lateral R@1m(%)	Longitudinal R@1m(%)	Yaw R@1°(%)
kitti_0_0	0	0	3.964	10.096	4.923	48.98	15.02	27.04
kitti_10_5	10	5	1.161	5.231	2.923	70.37	38.23	28.87
kitti_20_0	20	0	3.332	10.023	4.523	52.51	14.80	27.83
kitti_20_10	20	10	1.130	5.190	2.832	70.87	38.12	29.46
kitti_30_10	30	10	1.314	5.319	2.963	67.26	36.90	29.65

TABLE A.5: Performance under extreme test perturbations (Test: 30m, 10°)

Training Config	XY Pert. (m)	Yaw Pert. (°)	Lateral MAE(m)	Longitudinal MAE(m)	Yaw MAE(°)	Lateral R@1m(%)	Longitudinal R@1m(%)	Yaw R@1°(%)
kitti_0_0	0	0	6.596	15.766	5.941	43.01	12.44	24.61
kitti_10_5	10	5	1.470	7.767	3.059	68.77	36.00	28.60
kitti_20_0	20	0	5.778	15.562	5.299	47.11	12.94	25.48
kitti_20_10	20	10	1.482	7.512	2.990	69.16	35.99	28.98
kitti_30_10	30	10	1.821	7.879	3.171	65.50	34.59	29.05