MH-CDNet: Map- and History-Aided Change Detection of Traffic Signs in High-Definition Maps

Yangyi Zhong, Yuxiang Guo, Peng Yue, Chuanwei Cai, Jian Li, Kai Yan

***Abstract*— Recent high-definition maps play a key role in autonomous driving, providing precise environmental information for navigation and decision-making. Current approaches typically focus on either the online map construction or the timely updating of existing maps to keep them aligned with the real-world environment. However, the first approach is limited by insufficient map accuracy, while the second lacks emphasis on map currentness, especially for map change detection. Thus, we propose a novel method that integrates spatial-temporal information from both prior maps and revisit records to enable timely and accurate high-definition map change detection. The experimental results demonstrate that our method significantly outperforms traditional approaches for this task, achieving a 35% improvement in accuracy on real-world data. To the best of our knowledge, this is the first method for high-definition map change detection that combines both map-aided and history-aided components. It significantly enhances the model's ability to discern map changes in complex environments and partially addressing the limitations of previous approaches.**

I. Introduction

High-definition (HD) maps play a key role in autonomous driving systems by providing precise real-world information necessary for downstream tasks such as navigation and decision-making in complex road scenarios [1]. However, the real world is dynamic: Lane markings, traffic signs, and intersections change over time. As these changes accumulate, outdated HD maps may fail to accurately reflect the current road conditions, leading to erroneous judgments by autonomous vehicles and posing significant safety risks [2]. Conventionally, updating HD maps typically involves using specialized mapping tools to refresh the entire map as a whole. This approach is time-consuming and costly, making it difficult to keep the maps up-to-date with the latest environmental changes [3]. Therefore, efficiently obtaining HD maps that reflect the current real-world environment has become an urgent challenge.

One direct approach to ensuring that HD maps accurately reflect the most current real world is to use sensors (e.g., RGB cameras) equipped in vehicles for online HD map construction [4]. Prior research [5], [6] has explored various methods to achieve this goal. Furthermore, some studies [7], [8] have leveraged offline map to assist in this online mapping process, making certain progress. However, these methods are still struggling to achieve the precision and robustness required for deployment in real-world scenarios, indicating a gap between current capabilities and practical application needs [9], [10].

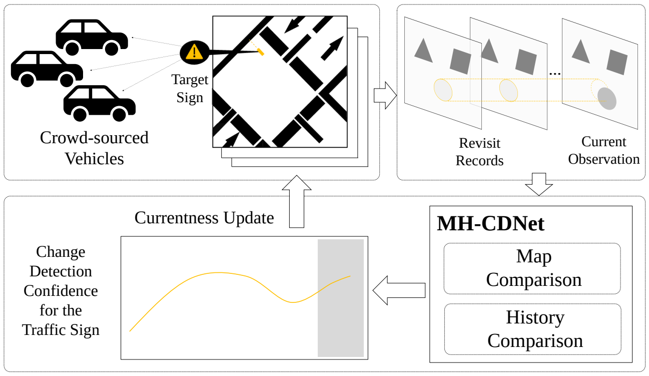

Fig. 1. Overview of the framework: Crowd-sourced vehicles are employed to detect traffic signs and obtain observations of the same traffic signs over different time periods. Each observation is processed by the MH-CDNet model, which compares the observation with prior maps and revisit records, outputting a change detection confidence score. This score is used to determine whether the traffic sign has changed and triggers an update to ensure the currentness of the sign.

Another research direction focuses on timely performing HD map change detection and updating only the changed sections, aiming to overcome the time-consuming and costly drawbacks of traditional HD map surveying. However, most existing studies [11], [12] primarily determine whether a map element has changed by comparing the vehicle's current perception data with the HD maps. Due to the inherent complexity of real-world environments, these methods lack robustness and reliability, as discrepancies between detected elements and map data often arise not from genuine changes but from temporary occlusions, sensor noise, or other transient factors [13], [14], [15]. Some studies [16], [17] use revisit data, consisting of repeated historical observations of the same map elements, to infer changes. By comparing current observations with these revisit records, they reduce errors caused by transient factors in single comparison. However, these methods often rely solely on detection confidence from object detection models at the time of observation, while overlooking other crucial information, such as the

This study was supported by the Chongqing Technology Innovation and Application Development Project [Grant No. CSTB2022TIAD-DEX0013]; funding from the State Key Laboratory of Intelligent Vehicle Safty Technology and Chongqing Changan Automobile Co. Ltd.

*Corresponding author: Yuxiang Guo

Yangyi Zhong, Yuxiang Guo, Peng Yue and Chuanwei Cai are with School of Remote Sensing Information Engineering, Wuhan University, Wuhan, China. (e-mail: yangyi.zhong@whu.edu.cn, whursgyx@whu.edu.cn, pyue@whu.edu.cn, chuaiwei.cai@whu.edu.cn).

Yangyi Zhong, Jian Li and Kai Yan are with State Key Laboratory of Intelligent Vehicle Safty Technology, Chongqing Changan Automobile Co.Ltd, Chongqing, China. (e-mail: yangyi.zhong@whu.edu.cn, Lijian1@changan.com.cn, yankai1@changan.com.cn).

Peng Yue is with Hubei Province Engineering Center for Intelligent Geoprocessing, Wuhan University, Wuhan, China. (e-mail: pyue@whu.edu.cn).

coordinates and image representation of the corresponding map element. This narrow focus neglects spatial relationships and detailed environmental context, making it difficult to accurately distinguish genuine map changes from false detections.

What's more, most prior works have relied on two types of datasets: custom-built datasets [18] collected using specialized high-precision surveying vehicles, and widely used public datasets [19], [2], [20] gathered through similar methods. Both approaches provide high accuracy but come with significant costs, limiting scalability and practical use [21].

To address the aforementioned shortcomings, we first propose the construction of a low-cost revisit dataset specifically designed for traffic sign change detection in HD maps. This dataset, collected through crowd-sourced vehicles and commercial sensors, provides an affordable and scalable solution for detecting traffic sign changes, eliminating the need for expensive surveying vehicles typically used for HD map updates. Building upon this dataset, we introduce a novel method for HD map change detection that integrates prior maps and revisit records. By comparing current observation with prior maps and revisit records, our method focuses on identifying and distinguishing genuine changes in traffic signs from false discrepancies. To validate the effectiveness of our method, we conduct comprehensive experiments using traffic signs as a case study. The experiments demonstrate a 35% improvement in overall detection success rate, showcasing the method's ability to accurately detect and classify changes.

In summary, our contributions include:

1) We introduce the first low-cost revisit dataset specifically designed for detecting traffic sign changes in HD maps, offering a valuable resource for advancing research in this area.
2) We propose the first method for HD map change detection that integrates prior maps and revisit records to identify changes. It can accurately distinguish genuine map changes from false discrepancies.
3) We validate the effectiveness of our method through extensive experiments, demonstrating its robustness in detecting map changes, with a notable 35% increase in overall detection success rate compared to traditional methods.

II. Related Work

A. Online HD Map Construction

Recent advancements have focused on using onboard sensor data to construct bird's-eye view (BEV) representations for HD mapping, offering a promising alternative to manual map creation [4]. HDMapNet [5] pioneered this approach, converting raster map segmentation into vector maps through post-processing, setting benchmarks for vector HD mapping. However, its reliance on post-processing limits real-time applicability. To address this, VectorMapNet [22] and MapTR [23] used DETR-based transformers[24] for direct end-to-end prediction, eliminating post-processing.

Some subsequent studies focus on improving the geometric modeling of map elements to enhance precision, such as BeMapNet (Bézier curves) [25], PivotNet (point-to-line) [26], and GeMap (G-Representation) [27]. Models like MapTRv2 [28] and HIMap [29] improve querying mechanisms, enhancing performance in complex scenarios. Temporal information is also crucial, with StreamMapNet [6] introducing temporal optimization for better adaptation to dynamic environments. Approaches like MapEX [7] and P-MapNet [8] have pioneered the use of prior maps to enhance the performance of online map construction. However, experiments by Bateman et al. show that incorporating prior maps can effectively handle minor adjustments, but struggles with larger changes [10].

While these methods show promise in addressing specific challenges, they still face inherent limitations for autonomous driving. Most existing methods are constrained by a narrow focus on a limited set of map objects, such as lane markings, with crucial elements like traffic signs often excluded. Additionally, the accuracy of these maps remains suboptimal, frequently falling behind offline maps that are processed in a more comprehensive and unified manner. As a result, these shortcomings limit the practical application in real-world scenarios, where more complete and accurate maps are required for reliable navigation and decision-making.

B. HD Map Change Detection

Advances in artificial intelligence and intelligent systems in mass-produced vehicles have rapidly progressed in recent years, making crowd-sourced map updates feasible — a promising approach that increases update frequency and significantly reduces costs [3].

Heo et al. used direct metric learning to project HD maps into image space to detect changes, comparing the projected maps with detected map elements and computed a similarity score [11]. Lambert et al. introduced a "Trust but Verify" (TbV) dataset and trained learning-based algorithms to detect changes in HD vector maps [2]. Park et al. used YOLOv3 to detect traffic signs, lights, and road markings, evaluating changes by comparing the center point coordinates of detected objects with their corresponding elements in HD maps [30]. Wild et al. proposed ExelMap, which leverages real-time 360° camera images compared with HD maps in BEV view, focusing on element-based, interpretable map change detection [12]. These methods provide effective solutions for instantaneous detection, but they do not account for temporal information, making them vulnerable to errors caused by transient factors such as temporary occlusions or momentary sensor noise. To address this limitation, Diff-Net leveraged ConvLSTM to incorporate information from consecutive frames [18]. However, it still relies on a single comparison to make decisions about changes. As a result, if transient factors persist across the observation period, the detection remains prone to errors.

Different from these single comparison methods, several approaches have adopted a multiple comparisons strategy, considering data from multiple observations over time.

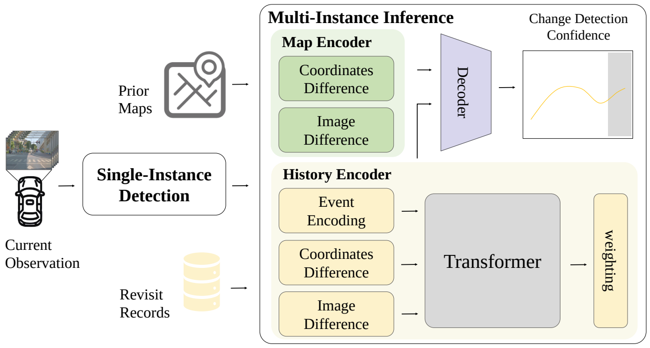

Fig. 2. Overview of the pipeline: In the single-instance detection stage, the current observation is processed to extract spatial-temporal features. In the multi-instance inference stage, these features are compared with prior maps and revisit records, which are processed through the **Map Encoder** and **History Encoder**, respectively. The outputs are then dynamically integrated by the **Decoder** to determine if a traffic sign has changed.

Benny et al. utilized a multi-instance framework to validate updates; however, their crowd-sourced method overlooked the timing and quality of individual detections, failing to account for the varying reliability of different observations [31]. Jo et al. applied Dempster-Shafer theory to assess detection map elements and measure map changes [17]. Tchuente et al. utilized Bayesian inference to determine the existence of map elements [16]. But traditional methods like these rely solely on detection confidence from object detection models, neglecting richer information available in revisit data. For example, if a map element is partially occluded or truly absent, both cases might result in low detection confidence, making the element undetectable; however, the image of a partially occluded element would more closely resemble the corresponding map, allowing the system to correctly identify that no genuine change has occurred.

In contrast, our method explicitly considers traffic sign changes. By performing timely change detection, it leverages the precision of offline maps while maintaining their currentness. Compared to existing HD map change detection methods, our method integrates consecutive frames to address transient factors in single-instance detection and combines revisit records based on their timestamps and data quality in multi-instance inference. By leveraging deep learning, it effectively extracts valuable insights from both map and revisit data, reducing the over-reliance on object detection confidence and enhancing map change detection accuracy and system reliability in real-world scenarios.

III. Method

Our method is based on a two-stage framework: single-instance detection and multi-instance inference. In the first stage, map elements are detected from current observation, and key spatial-temporal features—such as a traffic sign's position, appearance, and detection confidence—are extracted for next stage. Each observation is also stored as a revisit record to provide temporal context for future assessments. In the second stage, the extracted features are compared with prior maps and revisit records collected over multiple time periods. By integrating these comparisons, the method accurately identifies genuine map changes while filtering out false events caused by transient or misleading factors.

A. Problem formulation

Our method operates on three primary inputs: current observation, prior maps, and revisit records, with the objective of determining whether the currently observed traffic sign remains unchanged. In this context, the detection of the sign's existence is equivalent to identifying no change, while its absence indicates a change.

The current observation $\mathcal{N}$ consists of the observed coordinates $\mathbf{c}_n$, an associated image representation $\mathbf{i}_n$, and a timestamp $\mathbf{t}_n$. This data captures real-time observation of traffic signs from the autonomous vehicle's perception system.

$$\mathcal{N} = \{\mathbf{c}_n, \mathbf{i}_n, t_n\}, \tag{1}$$

The prior maps $\mathcal{M}$ contain the expected coordinates $\mathbf{c}_m$ of the corresponding traffic sign and its associated textual description $\mathbf{s}_m$.

$$\mathcal{M} = \{\mathbf{c}_m, \mathbf{s}_m\}, \tag{2}$$

The revisit records $\mathcal{R}$ contain a series of records of the current traffic sign. Each record, R_i is represented by coordinates $\mathbf{c}_{r_i}$, an associated image representation $\mathbf{i}_{r_i}$, an image quality index q_{r_i}, a timestamp t_{r_i}, and an event type e_{r_i}.

$$R_i = \{\mathbf{c}_{r_i}, \mathbf{i}_{r_i}, q_{r_i}, t_{r_i}, e_{r_i}\}, \tag{3}$$

$$\mathcal{R} = \{R_1, R_2, \ldots, R_i\}, \quad i = 1, 2, \ldots, N_{\max}, \tag{4}$$

where $N_{\max}$ is the maximum number of revisit records considered for each traffic sign.

Formally, we aim to determine the existence P_{exist} of a traffic sign as a function of these inputs:

$$P_{\text{exist}} = f(\mathcal{N}, \mathcal{M}, \mathcal{R}). \tag{5}$$

Here, the function $f(\cdot)$ represents our proposed method.

B. Single-Instance Detection

The single-instance detection stage involves two steps: target detection and coordinate calculation.

Target Detection: Traffic signs are detected from observation images using a pre-trained model, which outputs the bounding box coordinates, category, and confidence of each detected traffic sign. To improve localization robustness, a distinctive corner point near the bounding box center is identified using feature detection methods. The image with the highest detection confidence is cropped based on the selected corner point and stored as the representative image for subsequent multi-instance inference.

Coordinate Calculation: The coordinates of detected traffic signs in the real world are computed using a collinearity equation, linking the image positions (x, y) to the world coordinates (X, Y, Z) via camera parameters.

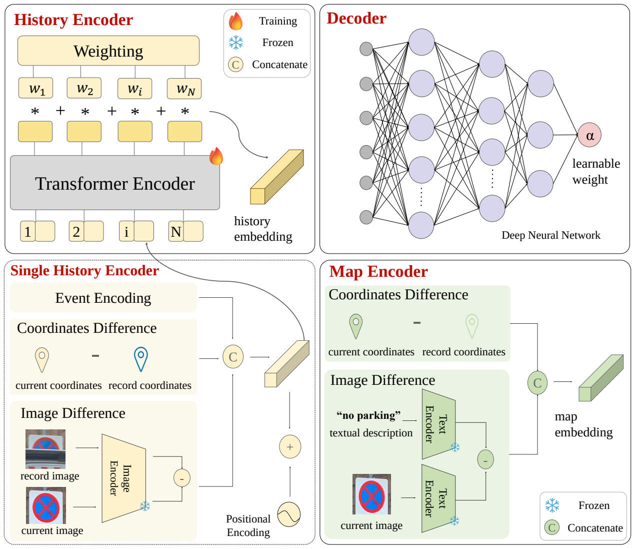

Fig. 3. Details of the multi-instance inference stage: The **Map Encoder** compares the current observation with map elements using coordinates and image differences. The **History Encoder** compares the current observation with revisit records. In the *Single History Encoder*, coordinates and image differences are computed for each revisit record, along with event type encoding. The Transformer is then employed to model temporal dependencies across revisit records. The **Weighting** module assigns weights based on record quality and time gaps. The outputs from the Transformer are then fused together based on these weights to create the final history representation. The **Decoder**, a deep neural network (DNN), learns adaptive weights to combine the features from both encoders, and fuses them to produce the final change detection result

$$\begin{bmatrix} x-x_0 \\ y-y_0 \\ -f \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} X-X_s \\ Y-Y_s \\ Z-Z_s \end{bmatrix}. \tag{6}$$

where x_0, y_0 represent the principal point offsets, f is the focal length of the camera, X_s, Y_s, Z_s denote the camera coordinates in the world coordinate system, r_{ij} are elements of the parameters matrix R that aligns the camera coordinate system to the world coordinate system.

Using data from multiple captured frames, a weighted least squares adjustment is applied to refine the coordinate estimates:

$$\hat{\mathbf{X}} = (\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{W}\mathbf{b}. \tag{7}$$

where $\mathbf{L}$ represents the observation matrix derived from the collinearity equations, $\mathbf{W}$ is a diagonal weight matrix based on detection confidences, and $\mathbf{b}$ contains the observed values.

C. Multi-Instance Inference

The multi-instance inference stage processes single-instance detection outputs by comparing them separately with prior maps through the Map Encoder and with revisit records through the History Encoder. These comparison results are then fused in the Decoder to produce a final change assessment.

1) Map Encoder: This module compares the current observation with the corresponding map elements to evaluate their similarity. In cases where the quality of revisit records is poor, this comparison enhances the robustness of inference.

Coordinates Difference: The coordinates of the current observation traffic sign are compared with the coordinates of the corresponding traffic sign in the HD map.

$$\Delta\mathbf{c} = \phi(\mathbf{c}_n - \mathbf{c}_m, \mathbf{W}_c), \tag{8}$$

where $\phi(\cdot,\cdot)$ represents the linear transformation with a learnable parameter matrix $\mathbf{W}_c$.

Image Difference: The image $\mathbf{i}_n$ associated with the current observation and the textual description $\mathbf{s}_m$ of the corresponding traffic sign in the HD map are processed separately using a CLIP-based [32] encoder, resulting in image features $\mathbf{i}'_n$ and text features $\mathbf{s}'_m$. The difference is computed as:

$$\Delta\mathbf{i} = \phi(\mathbf{i}'_n - \mathbf{s}'_m, \mathbf{W}_i), \tag{9}$$

where $\phi(\cdot,\cdot)$ represents the linear transformation with a learnable parameter matrix $\mathbf{W}_i$.

Fusion: The coordinates and image difference embeddings are concatenated, which integrates both spatial and image feature information.

$$\Delta Map = [\Delta\mathbf{c}, \Delta\mathbf{i}]. \tag{10}$$

2) History Encoder: This module compares the current observation with revisit records to assess similarity. A maximum number of records is selected based on their timestamps, and each record is processed in the Single History Encoder to compute coordinates and image differences and encode event types. These features are then integrated for temporal analysis through a Transformer [33] to capture dependencies.

Coordinates Difference: The coordinates of the current observation traffic sign are compared with the corresponding coordinates of each record r_i.

$$\Delta\mathbf{c}_{r_i} = \phi(\mathbf{c}_n - \mathbf{c}_{r_i}, \mathbf{W}_c), \tag{11}$$

where $\phi(\cdot,\cdot)$ represents the linear transformation with learnable parameter matrix $\mathbf{W}_c$.

Image Difference: Similarly, the current image $\mathbf{i}_n$ and each record image $\mathbf{i}_{r_i}$ are processed by the same image encoder to obtain the corresponding image features$\mathbf{i}'_n$, $\mathbf{i}'_{r_i}$. The difference is computed as:

$$\Delta\mathbf{i}_{r_i} = \phi(\mathbf{i}'_n - \mathbf{i}'_{r_i}, \mathbf{W}_i), \tag{12}$$

where $\phi(\cdot,\cdot)$ represents the linear transformation with learnable parameter matrix $\mathbf{W}_i$.

Event Encoding: Each record is assigned an event type reflecting how the single-instance detection outcome aligns with the prediction existence in the multi-instance inference stage. A True Positive signifies that a detected sign is indeed present in the real world, while a True Negative indicates the absence of both a detected sign and an actual sign. A False Positive occurs when the system detects a sign at a location where none exists, and a False Negative arises when no sign is detected despite its actual existence. By encoding these event types in a one-hot vector, the method captures the extent to which each individual observation corresponds to physical reality, thereby enabling more reliable integration of historical data.

Temporal Attention: These embeddings are concatenated to form a combined embedding for each record $\mathbf{E}_{\text{re\_c}_i}$:

$$\mathbf{E}_{\text{re\_c}_i} = \text{concat}(\Delta\mathbf{c}_{r_i}, \Delta\mathbf{i}_{r_i}, \mathbf{d}_{r_i}), \tag{13}$$

The revisit record embedding, along with its corresponding positional embedding $\mathbf{p}_{r_i}$, is then passed through a Transformer-based encoder to capture temporal dependencies and contextual relevance. The final input to the Transformer is:

$$\mathbf{E}_{\text{input},i} = \mathbf{E}_{\text{re\_c},i} + \mathbf{p}_{r_i}, \quad i = 1, 2, \ldots, N_{\max}, \tag{14}$$

where $N_{\max}$ represents the maximum number of revisit records considered.

The model processes this input using the Transformer, considering temporal dependencies, to generate the output:

$$\mathbf{E}_{\text{output}} = \text{Transformer}(\mathbf{E}_{\text{input}}), \tag{15}$$

Weighting: To evaluate the relevance and reliability of each record, we compute the weight w_{r_i} based on the time difference Δt_{r_i} and the image quality index q_{r_i}. The weight function is given by:

$$\Delta\mathbf{t}_{r_i} = \mathbf{t}_n - \mathbf{t}_{r_i}, \tag{16}$$

$$w_{r_i} = \frac{(Q_{\max} - q_{r_i})}{\Delta t_{r_i} + 1}. \tag{17}$$

where $Q_{\max}$ is the maximum image quality value.

Fusion: The final history representation is obtained by applying the computed weights w_{r_i} to the outputs of the Transformer:

$$\Delta History = \sum_i w_{r_i} \cdot \mathbf{E}_{\text{output},i}, \quad i = 1, 2, \ldots, N_{\max}. \tag{18}$$

This weighted summation integrates the contributions of all revisit records, providing better differentiation between genuine changes and false changes compared to single comparison.

3) Decoder: This module fuses the outputs from the Map Encoder and the History Encoder to produce a final prediction by leveraging a adaptive fusion mechanism.

Adaptive Fusion: A learnable weight α modulates the contributions of each comparison based on their conditions. This mechanism, optimized during training, effectively accounts for data quality and recency, enhancing prediction accuracy and robustness.

$$\alpha = \sigma(\mathbf{W}_\alpha \cdot \text{concat}(\Delta Map, \Delta History)), \tag{19}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\mathbf{W}_\alpha$ represents learnable parameters. The final output is computed as a weighted sum of the two predictions:

$$\text{output} = \alpha \cdot \Delta Map + (1 - \alpha) \cdot \Delta History. \tag{20}$$

Sign Loss: The output is then compared with the ground truth label $y \in \{0, 1\}$, indicating whether the traffic sign exists, using the Binary Cross-Entropy (BCE) loss function.

$$\mathcal{L}_{\text{BCE}} = -\left(y \cdot \log(\text{output}) + (1 - y) \cdot \log(1 - \text{output})\right). \tag{21}$$

Minimizing this loss function enables the model to accurately predict the existence of traffic signs by optimizing both the individual components and the fusion mechanism.

IV. Experiments

In order to evaluate the performance of our method, we conduct experiments using data collected from consumer-grade vehicles in real-world scenarios. Compared to existing methods [17], [16] that leverage revisit records, our method excels in accurately discerning genuine map changes, achieving significant improvements over current state-of-the-art methods.

A. Data Collection

Our experimental vehicle was equipped with an 8MP 60fps camera module and an advanced integrated navigation system (INS) for precise positioning. The INS achieved a positioning accuracy of $2\,\text{cm} + 1\,\text{ppm}$ in RTK mode, which is critical for applications requiring centimeter-level precision. Additionally, it provided attitude accuracy within 0.1° (1σ) for roll and pitch, ensuring reliable attitude estimation essential for stable platform control and accurate data collection. The crowd-sourced vehicles, equipped with these sensors, collected data across various routes, forming an extensive dataset.

The routes spanned approximately 117 kilometers, encompassing diverse environments such as residential streets, arterial roads, and intersections. Data collection was performed at various times of the day and under different weather conditions, incorporating natural variability.

Additionally, challenges such as sensor noise and motion blur were present in the collected data. To further enrich the dataset and enhance its diversity, manual data augmentation was applied to simulate various scenarios.

Within these routes, a total of **288 traffic signs** were observed, encompassing **46 distinct categories** such as speed limits, stop signs, and pedestrian crossings. Each traffic sign was revisited multiple times. This crowd-sourced revisit strategy ensures sufficient temporal coverage to validate changes in map elements.

The dataset encompasses diverse event types, including **true positives (TP)**, **false positives (FP)**, **true negatives (TN)**, and **false negatives (FN)**, providing comprehensive data to evaluate model performance. Table I summarizes the distribution of these event types.

To address the scarcity of real-world map changes, such as disappearing signs, we simulate these cases using Edge-Connect, a method based on Variational Autoencoder (VAE) [35]. This approach generates realistic imagery of missing signs, ensuring closer alignment with real-world conditions compared to traditional masking methods, thereby enriching the dataset's diversity and authenticity.

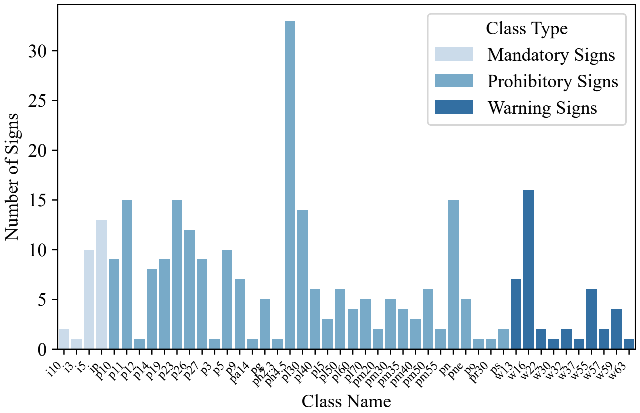

Fig. 4. Distribution of sign instances across traffic sign classes. The traffic sign categories used in this study are based on the naming conventions derived from the TT100K dataset [34].

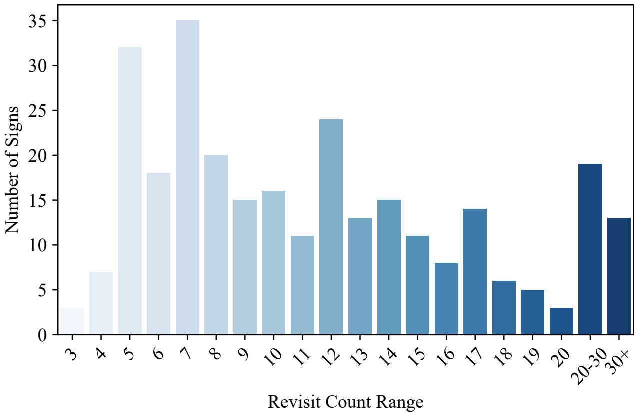

Fig. 5. Distribution of revisit counts across traffic signs.

B. Implementation Details

The implementation of our method was conducted on a GeForce RTX 3090 GPU, with a batch size of 64 and an initial learning rate of 1×10^{-5}, using the Adam optimizer with a weight decay of 1×10^{-5}. Training was performed over 50 epochs.

For object detection, we utilized YOLOv10 [36] which fine-tuned on the TT100K dataset for 30 epochs. For aligning image and text representations of traffic signs, CLIP was similarly fine-tuned on the TT100K dataset. For determining the traffic sign's location in the image, Harris Corner Detection [37] was employed to identify a stable point near the bounding box center. For capturing temporal dependencies, we configured the Transformer encoder with 4 layers, 8 attention heads. For evaluating revisit records, the BRISQUE metric [38] was used to assess image quality, guiding the weighting mechanism accordingly. The maximum number of revisit records per traffic sign was set to 3, balancing computational efficiency with the need to capture meaningful temporal patterns.

To validate the effectiveness of our method, we conducted comparative experiments against existing change detection methods that utilize revisit data. These methods are based on Bayesian inference and Dempster-Shafer theory. We applied these approaches, as well as our own, to the same dataset in order to demonstrate how our method improves upon the handling of dynamic changes and data inconsistencies.

TABLE I
DISTRIBUTION OF EVENT TYPES IN THE DATASET

Event Type	**Count**	**Percentage (%)**
True Positives (TP)	1754	49.16
False Positives (FP)	812	22.76
False Negatives (NP)	605	16.95
True Negatives (TN)	397	11.13

C. Results on Collected Datasets

Unlike traditional object detection methods that rely on standard metrics such as precision, recall, and F1-score, our model focuses on identifying and correcting misdetections. For instance, in the case of a FP event, even though the traffic sign is detected in the single-instance detection phase, we classify the observation as ”absent” based on the multi-instance inference. This can be considered a successful detection for this event, as it correctly identifies the absence of the sign in the environment. Therefore, our evaluation metrics are based on the success rate of correctly identifying the different event types.

$$S_{\text{event type}} = \frac{1}{N_{\text{event type}}} \sum_{i=1}^{N_{\text{event type}}} \mathbb{I}(\hat{y}_i = y_i), \tag{22}$$

$$S_{\text{overall}} = \sum_{j} p_j \cdot S_{\text{event type}_j}. \tag{23}$$

where $\hat{y}_i$ represents the predicted label for the i-th record, and y_i is the ground truth label. $\mathbb{I}(\cdot)$ is an indicator function, which returns 1 when the prediction is correct, and 0 otherwise. $S_{\text{event type}_j}$ is the success rate for event type j, p_j is the percentage of event type j in the dataset, and the sum is taken over all event types.

To assess our method's effectiveness, we compared it with similar techniques that use revisit data for HD map change detection. The comparative results, as presented in Table II, highlight the overall superiority of our method, showing an approximate 35% improvement. This superiority is reflected in our method's ability to discern genuine map changes, confirming its robustness and reliability in dynamic environments.

Notably, our approach excels in addressing FN events. Traditional methods, which rely heavily on detection confidence from single-instance detection, struggle to differentiate temporary factors such as occlusions or noise from genuine map changes, often resulting in misclassifications. In contrast, our method distinguishes transient factors from actual changes, ensuring greater reliability in real-world scenarios.

Meanwhile, the slightly lower success rate for TN and FP events in our method can also be attributed to the higher reliance of traditional methods on detection confidence. Since these cases often involve low detection confidence,

TABLE II

COMPARISON OF METHODS

Method	Overall(%)	TP (%)	TN (%)	FP (%)	FN(%)
Bayesian	45.43	24.39	100.00	87.50	2.17
DS	43.90	25.00	84.21	82.50	10.87
Ours	78.96	92.68	71.05	57.50	73.91

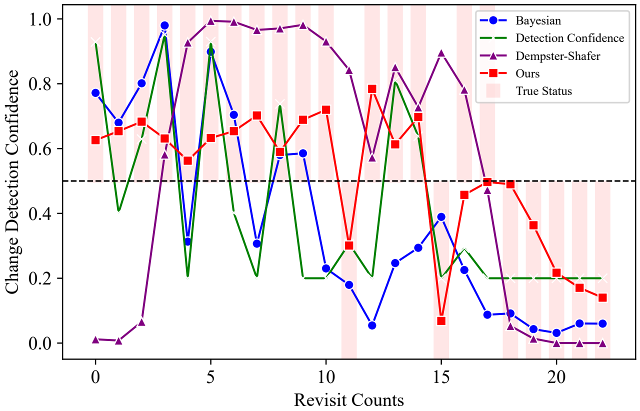

Fig. 6. Comparison of Predicted Change Detection Confidence for the Same Traffic Sign Across Different Revisit Counts. In each revisit event, the light red shaded area indicates the **true status** of the traffic sign: above 0.5 is considered present, below 0.5 is considered absent. A prediction is considered successful when it falls within this designated area.

TABLE III

ABLATION STUDY RESULTS

Setting	Overall(%)	TP(%)	TN(%)	FP(%)	FN(%)
Map Only	73.17	90.85	55.26	42.50	78.26
History Only	77.44	85.37	76.32	55.00	89.13
No Weighting - H	74.09	91.46	60.53	33.75	93.48
No Weighting - D	58.84	53.05	86.84	68.75	39.13
Full Model	78.96	92.68	71.05	57.50	73.91

traditional methods tend to mirror this decrease, unintentionally leading to lower prediction probabilities and higher success rates by default. However, these methods merely track fluctuations in detection confidence without considering the actual presence or absence of the traffic sign. In contrast, our approach maintains a balanced and reliable performance across all event types.

In Fig. 6, we illustrate the predicted change detection confidence for a traffic sign across multiple revisit records using different methods. Bayesian method closely follows the detection confidence, as reflected in its erratic and fluctuating behavior. Similarly, Dempster-Shafer (DS) method demonstrates greater variability in predictions, with inconsistent results across all event types, making it prone to misclassifications in dynamic scenarios. In contrast, our method's predictions are closest to the true status, while showing a smoother and more robust trend, effectively correcting misdetections even in challenging scenarios such as occlusions.

D. Ablation Studies

To evaluate the contributions of different components in our method, we conducted a series of ablation studies. These experiments tested the performance of using prior maps only (*Map Only*), revisit records only (*History Only*), and configurations without weighting fusion (*No Weighting - H* for history encoder and *No Weighting - D* for decoder). The results are presented in Table III.

Map Only: Using prior maps only, the model achieves a little higher FN success rate, as it relies solely on the spatial precision of map to infer the absence of traffic signs. However, the lack of temporal context limits its ability to handle transient changes, resulting in reduced performance on other metrics like TN and FP. This highlights the importance of integrating temporal information for robust performance in dynamic environments.

History Only: When relying solely on revisit records, the model shows improved performance in identifying negative events (TN and FN). This is because temporal information helps distinguish transient factors from permanent changes. However, without the spatial precision provided by map data, the model exhibits reduced success rates for TP and FP, particularly when historical data quality is low or outdated.

No Weighting - H: In the **History Encoder**, without the weighting mechanism, all revisit records are treated equally, regardless of their quality or recency. This configuration occasionally improves FN success rates, as it avoids over-penalizing older or lower-quality observations. However, the lack of prioritization for high-quality and recent events leads to a decrease in overall robustness, as evidenced by other lower success rates.

No Weighting - D: In the **Decoder**, without adaptive weighting, the decoder simply averages the two input sources, prior maps and revisit records, by dividing them equally. This fixed fusion method fails to adjust to the varying relevance of each source in different contexts. As shown in the experimental results, both TP and FN represent instances where traffic signs should be present, but the accuracy for these cases is low. It indicates that the model tends to favor the absence of traffic signs.

Full Model: Although certain ablated configurations achieve higher success rates on specific metrics, these gains often come at the expense of reduced performance in other event types. The full model effectively balances the strengths of different components, delivering consistent and robust performance across all metrics, making it well-suited for diverse real-world scenarios.

V. CONCLUSION

In this work, with both map-aided and history-aided components, we introduced a novel method for HD map change detection. By comparing current observation with the prior maps and revisit records, our method effectively handles the complexities of real-world environments to make accurate judgments. Extensive experiments validate the effectiveness of our method, highlighting its superiority over traditional methods in accurately distinguishing genuine map changes from false discrepancies.

As a future direction, we plan to extend this method to other map elements, such as lane markings and road boundaries, broadening its applicability for scalable and efficient map maintenance in autonomous driving systems.

References

[1] K. Wong, Y. Gu, and S. Kamijo, “Mapping for autonomous driving: Opportunities and challenges,” *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 1, pp. 91–106, 2021.

[2] J. Lambert and J. Hays, “Trust, but verify: Cross-modality fusion for hd map change detection,” 2022. [Online]. Available: https://arxiv.org/abs/2212.07312

[3] B. Wijaya, K. Jiang, M. Yang, T. Wen, Y. Wang, X. Tang, Z. Fu, T. Zhou, and D. Yang, “High definition map mapping and update: A general overview and future directions,” 2024. [Online]. Available: https://arxiv.org/abs/2409.09726

[4] X. Tang, K. Jiang, M. Yang, Z. Liu, P. Jia, B. Wijaya, T. Wen, L. Cui, and D. Yang, “High-definition maps construction based on visual sensor: A comprehensive survey,” *IEEE Transactions on Intelligent Vehicles*, pp. 1–23, 2023.

[5] Q. Li, Y. Wang, Y. Wang, and H. Zhao, “Hdmapnet: An online hd map construction and evaluation framework,” 2022. [Online]. Available: https://arxiv.org/abs/2107.06307

[6] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, “Streammapnet: Streaming mapping network for vectorized online hd map construction,” 2023. [Online]. Available: https://arxiv.org/abs/2308.12570

[7] R. Sun, L. Yang, D. Lingrand, and F. Precioso, “Mind the map! accounting for existing map information when estimating online hdmaps from sensor,” 2024. [Online]. Available: https://arxiv.org/abs/2311.10517

[8] Z. Jiang, Z. Zhu, P. Li, H. ang Gao, T. Yuan, Y. Shi, H. Zhao, and H. Zhao, “P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors,” 2024. [Online]. Available: https://arxiv.org/abs/2403.10521

[9] Z. Xie, Z. Pang, and Y.-X. Wang, “Mv-map: Offboard hd-map generation with multi-view consistency,” 2023. [Online]. Available: https://arxiv.org/abs/2305.08851

[10] S. M. Bateman, N. Xu, H. C. Zhao, Y. B. Shalom, V. Gong, G. Long, and W. Maddern, “Exploring real world map change generalization of prior-informed hd map prediction models,” 2024. [Online]. Available: https://arxiv.org/abs/2406.01961

[11] M. Heo, J. Kim, and S. Kim, “Hd map change detection with cross-domain deep metric learning,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 218–10 224.

[12] L. Wild, L. Ericson, R. Valencia, and P. Jensfelt, “Exelmap: Explainable element-based hd-map change detection and update,” 2024. [Online]. Available: https://arxiv.org/abs/2409.10178

[13] M.-Y. Yu, R. Vasudevan, and M. Johnson-Roberson, “Occlusion-aware risk assessment for autonomous driving in urban environments,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2235–2241, 2019.

[14] F. van Wyk, Y. Wang, A. Khojandi, and N. Masoud, “Real-time sensor anomaly detection and identification in automated vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1264–1276, 2020.

[15] L. Huang, Y. Zeng, S. Wang, R. Wen, and X. Huang, “Temporal-based multi-sensor fusion for 3d perception in automated driving system,” *IEEE Access*, vol. 12, pp. 119 856–119 867, 2024.

[16] D. Tchuente, D. Senninger, H. Pietsch, and D. Gasdzik, “Providing more regular road signs infrastructure updates for connected driving: A crowdsourced approach with clustering and confidence level,” *Decision Support Systems*, vol. 141, p. 113443, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923620301986

[17] K. Jo, C. Kim, and M. Sunwoo, “Simultaneous localization and map change update for the high definition map-based autonomous driving car,” *Sensors*, vol. 18, no. 9, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/9/3145

[18] L. He, S. Jiang, X. Liang, N. Wang, and S. Song, “Diff-net: Image feature difference based high-definition map change detection for autonomous driving,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2635–2641.

[19] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, “Argoverse 2: Next generation datasets for self-driving perception and forecasting,” 2023. [Online]. Available: https://arxiv.org/abs/2301.00493

[21] A. Das, J. IJsselmuiden, and G. Dubbelman, “Pose-graph based crowdsourced mapping framework,” in *2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*, 2020, pp. 1–7.

[22] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, “Vectormapnet: End-to-end vectorized hd map learning,” 2023. [Online]. Available: https://arxiv.org/abs/2206.08920

[23] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, “Maptr: Structured modeling and learning for online vectorized hd map construction,” 2023. [Online]. Available: https://arxiv.org/abs/2208.14437

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[25] L. Qiao, W. Ding, X. Qiu, and C. 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 (CVPR)*, June 2023, pp. 13 218–13 228.

[26] W. Ding, L. Qiao, X. Qiu, and C. Zhang, “Pivotnet: Vectorized pivot learning for end-to-end hd map construction,” 2023. [Online]. Available: https://arxiv.org/abs/2308.16477

[27] Z. Zhang, Y. Zhang, X. Ding, F. Jin, and X. Yue, “Online vectorized hd map construction using geometry,” 2024. [Online]. Available: https://arxiv.org/abs/2312.03341

[28] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, “Maptrv2: An end-to-end framework for online vectorized hd map construction,” 2023. [Online]. Available: https://arxiv.org/abs/2308.05736

[29] Y. Zhou, H. Zhang, J. Yu, Y. Yang, S. Jung, S.-I. Park, and B. Yoo, “Himap: Hybrid representation learning for end-to-end vectorized hd map construction,” 2024. [Online]. Available: https://arxiv.org/abs/2403.08639

[30] Y.-K. Park, H. Park, Y.-S. Woo, I.-G. Choi, and S.-S. Han, “Traffic landmark matching framework for hd-map update: Dataset training case study,” *Electronics*, vol. 11, no. 6, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/6/863

[31] B. Wijaya, M. Yang, T. Wen, K. Jiang, Y. Wang, Z. Fu, X. Tang, D. O. Sigomo, J. Miao, and D. Yang, “Multi-session high-definition map-monitoring system for map update,” *ISPRS International Journal of Geo-Information*, vol. 13, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2220-9964/13/1/6

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[34] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, “Traffic-sign detection and classification in the wild,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, “Edgeconnect: Generative image inpainting with adversarial edge learning,” 2019. [Online]. Available: https://arxiv.org/abs/1901.00212

[36] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, “Yolov10: Real-time end-to-end object detection,” 2024. [Online]. Available: https://arxiv.org/abs/2405.14458

[37] C. G. Harris and M. J. Stephens, “A combined corner and edge detector,” in *Alvey Vision Conference*, 1988. [Online]. Available: https://api.semanticscholar.org/CorpusID:1694378

[38] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.