What Really Matters for Robust Multi-Sensor HD Map Construction?

Xiaoshuai Hao[1], Lingyu Liu[2], Yuting Zhao[3], Yuheng Ji[3], Luanyuan Dai[4], Shuai Cheng[5], Rong Yin[6]

***Abstract*— High-definition (HD) map construction methods are crucial for providing precise and comprehensive static environmental information, which is essential for autonomous driving systems. While Camera-LiDAR fusion techniques have shown promising results by integrating data from both modalities, existing approaches primarily focus on improving model accuracy, often neglecting the robustness of perception models—a critical aspect for real-world applications. In this paper, we explore strategies to enhance the robustness of multi-modal fusion methods for HD map construction while maintaining high accuracy. We propose three key components: data augmentation, a novel multi-modal fusion module, and a modality dropout training strategy. These components are evaluated on a challenging dataset containing 13 types of multi-sensor corruption. Experimental results demonstrate that our proposed modules significantly enhance the robustness of baseline methods. Furthermore, our approach achieves state-of-the-art performance on the clean validation set of the NuScenes dataset. Our findings provide valuable insights for developing more robust and reliable HD map construction models, advancing their applicability in real-world autonomous driving scenarios.**

I. Introduction

High-definition (HD) map construction is a critical task for autonomous driving systems, providing rich semantic and geometric road information essential for localization, perception, and path planning. HD maps capture key details such as lane boundaries and road markings, which are vital for the precise operation of autonomous vehicles. While most existing research focuses on improving the accuracy of HD map construction, multi-modal fusion approaches—integrating data from complementary sensors like cameras and LiDAR—have shown promising results by leveraging the strengths of both.

However, in real-world autonomous driving scenarios, perception systems must operate under diverse and often challenging conditions. These include sensor corruptions caused by adverse weather (*e.g.*, snow, fog), sensor failures (*e.g.*, camera crashes, LiDAR misalignment), and external disturbances, all of which can significantly degrade model performance. Despite these challenges, the robustness of HD map construction models—defined as their ability to sustain performance under such corruptions—has largely been overlooked in previous studies. This oversight creates a significant gap in ensuring the reliability and safety of autonomous driving systems.

[1]Beijing Academy of Artificial Intelligence. E-mail:xshao@baai.ac.cn.
[2]Baidu Inc.
[3]Institute of Automation, Chinese Academy of Science.
[4]Nanjing University of Science and Technology.
[5]China North Artificial Intelligent & Innovation Research Institute.
[6]Institute of Information Engineering, Chinese Academy of Sciences.

To address this gap, we investigate the robustness of multi-modal fusion methods for HD map construction while maintaining high accuracy. Specifically, we aim to answer two key questions: How do HD map construction models perform under various sensor corruptions, and what strategies can enhance their robustness without compromising accuracy? To achieve this, we propose three key components: data augmentation, a multi-modal fusion module, and training strategies. These components are designed to improve the resilience of HD map construction models against 13 types of multi-sensor corruptions, including both single-source and multi-source disruptions, as illustrated in Fig. 1.

We evaluate our approach on a *Multi-Sensor Corruption* dataset and benchmark its performance against baseline methods. Experimental results demonstrate that the proposed components significantly enhance the robustness of HD map construction models while achieving state-of-the-art performance on the clean validation set of the NuScenes dataset. These findings provide valuable insights for improving the robustness and reliability of HD map construction models, advancing their applicability in real-world autonomous driving systems. To summarize, the contributions of this paper are three-fold:

- **Comprehensive Robustness Benchmarking**: We conduct a systematic evaluation of multi-modal HD map construction methods using a dataset with 13 types of *Multi-Sensor Corruption*. This provides a thorough analysis of model performance under challenging conditions.
- **Enhancing Framework**: We propose three key components—data augmentation, a novel multi-modal fusion module, and a modality dropout training strategy—that significantly enhance the robustness of multi-modal fusion methods without sacrificing accuracy.
- **State-of-the-Art Performance**: Our approach not only strengthens model resilience against sensor corruptions but also achieves state-of-the-art results on the NuScenes dataset's clean validation set, demonstrating its effectiveness in real-world autonomous driving scenarios.

II. Related Work

A. HD Map Construction

The HD map construction task [1], [2], [3] focuses on generating high-resolution, precise maps that provide instance-level vectorized representations of geometric and semantic elements, such as lane boundaries and road structures. These maps are essential for accurate localization and path planning in autonomous driving systems. Recent advancements in camera-LiDAR fusion methods [4], [5], [6], [7], [8], [9] have

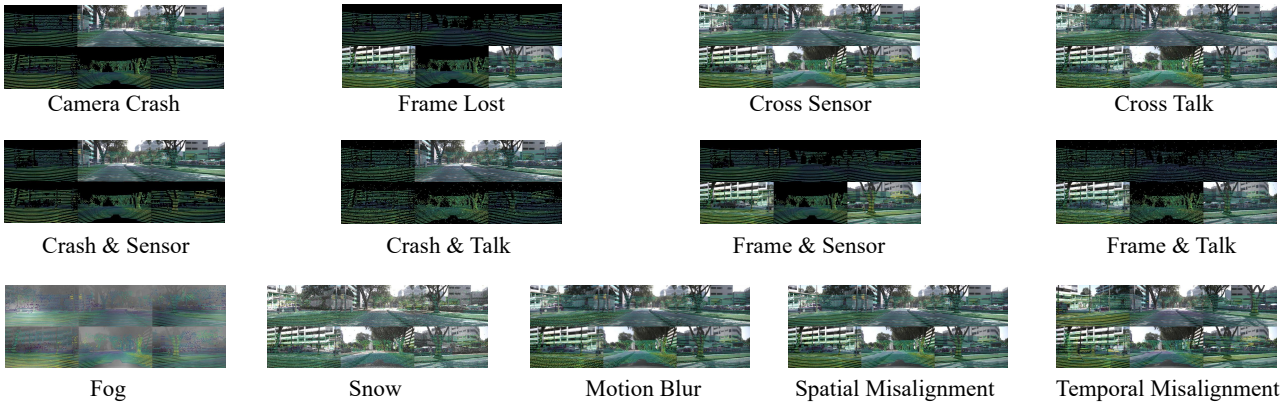

Fig. 1. **Overview of the Multi-Sensor Corruption dataset.** Multi-Sensor Corruption includes 13 types of synthetic camera-LiDAR corruption combinations that perturb both camera and LiDAR inputs, either separately or concurrently.

demonstrated the benefits of combining the semantic richness of camera data with the geometric precision of LiDAR. A particularly promising approach is BEV-level fusion, which encodes raw inputs from both sensors into a unified Bird's Eye View (BEV) space. This method effectively integrates complementary features from multiple modalities, achieving superior performance compared to uni-modal approaches. However, existing methods assume ideal conditions with complete and uncorrupted sensor data, leading to poor robustness in real-world scenarios where data may be missing or compromised. Such reliance on perfect sensor inputs often results in significant performance degradation or complete system failure under adverse conditions. In this paper, we investigate the critical factors necessary for achieving robust multi-sensor HD map construction.

B. Autonomous Driving Perception Robustness

Recent research has increasingly focused on the robustness of autonomous driving perception tasks [10], [11], [12], [13]. Studies such as RoBoBEV [14] evaluate the robustness of Bird's Eye View (BEV) perception, while others develop more resilient models or propose strategies to enhance system robustness [15]. Robo3D [16] benchmarks LiDAR-based semantic segmentation and 3D object detection under conditions of sensor corruption and failure. Zhu et al. [17] assess the natural and adversarial robustness of BEV-based models, introducing a 3D-consistent patch attack to improve spatiotemporal realism in autonomous driving. Additionally, MapBench [18] provides benchmarks for evaluating the robustness of HD map construction methods. In this paper, we investigate the robustness of camera-LiDAR fusion models for HD map construction by designing 13 types of corruption combinations that perturb camera and LiDAR inputs, either individually or simultaneously. Our proposed RoboMap model demonstrates superior robustness across diverse sensor failure scenarios. To the best of our knowledge, RoboMap is the first study to systematically explore the robustness of HD map construction under multi-sensor corruptions.

III. Multi-Sensor Corruption Dataset

Dataset Construction. In this paper, we investigate the robustness of camera-LiDAR fusion-based HD map construction tasks under various multi-sensor corruptions. Following the protocol established in [18], [19], we consider three corruption severity levels: *Easy*, *Moderate*, and *Hard*, for each type of corruption. The Multi-Sensor Corruption dataset is constructed by corrupting the *validation* set of the nuScenes dataset [20], which is widely adopted in recent HD map construction research. The Multi-Sensor Corruption dataset includes 13 types of synthetic camera-LiDAR corruption combinations, perturbing camera and LiDAR inputs either separately or concurrently, as illustrated in Fig. 1. These corruptions are categorized into three groups: camera-only, LiDAR-only, and multi-modal corruptions, addressing a wide range of real-world scenarios. Specifically:

- **Camera-Only Corruptions**: We design 2 types of corruptions using clean LiDAR data to simulate scenarios where the camera system is compromised while the LiDAR remains functional. These include:
  - *Camera Crash*: Simulates a complete failure of the camera system, where no visual data is available. This tests the model's ability to rely on LiDAR inputs.
  - *Frame Lost*: Mimics intermittent camera failures, where certain frames are dropped or missing. This evaluates the model's robustness to visual data.
- **LiDAR-Only Corruptions**: We create 2 types of corruptions using clean camera data to simulate scenarios where the LiDAR system is compromised while the camera remains operational. These include:
  - *Crosstalk*: Simulates interference between LiDAR sensors, where signals from one sensor affect another, leading to noisy or inaccurate point cloud data.
  - *Cross-Sensor*: Mimics misalignment or calibration errors between LiDAR sensors, resulting in inconsistent or distorted point cloud representations.
- **Multi-Modal Corruptions**: We propose 9 types of corruptions that perturb both camera and LiDAR inputs to simulate real-world scenarios where both modalities are affected. These include:
  - 4 combinations of the aforementioned failure types (*e.g.*, simultaneous *Camera Crash* and *Crosstalk*), testing the model's resilience to sensor failures.

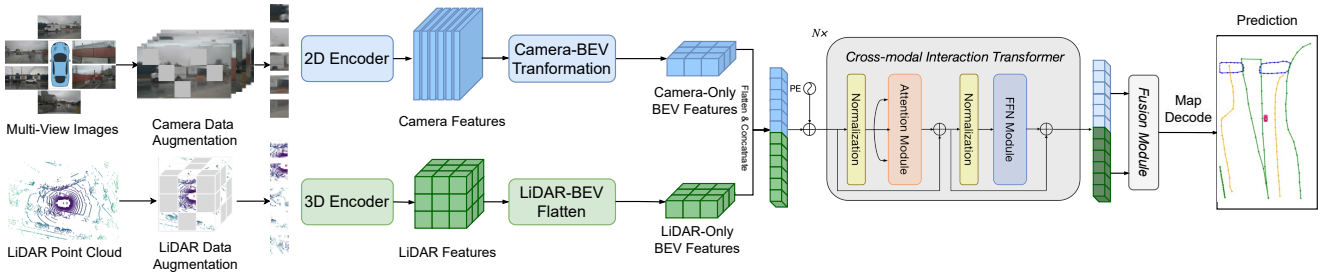

Fig. 2. **Overview of the RoboMap Framework.** The RoboMap framework begins by applying data augmentation to both camera images and LiDAR point clouds. Next, features are efficiently extracted from the multi-modal sensor inputs and transformed into a unified Bird's-Eye View (BEV) space using view transformation techniques. We then introduce a novel multi-modal BEV fusion module to effectively integrate features from both modalities. Finally, the fused BEV features are passed through a shared decoder and prediction heads to generate high-definition (HD) maps.

- 5 additional corruptions:
  - *Fog*: Simulates reduced visibility in both camera images and LiDAR point clouds due to dense fog.
  - *Snow*: Mimics the impact of snowfall, which can obscure camera images and scatter LiDAR signals.
  - *Motion Blur*: Represents blurring in camera images and distortion in LiDAR data caused by rapid vehicle motion.
  - *Spatial Misalignment*: Simulates misalignment between camera and LiDAR data due to calibration errors or physical shifts.
  - *Temporal Misalignment*: Mimics timing discrepancies between camera and LiDAR data, where inputs from the modalities are not synchronized.

Using this dataset, we conduct a systematic evaluation of the robustness of multi-modal HD map construction methods, providing a comprehensive analysis of model performance under adverse conditions.

Robustness Evaluation Metrics To assess the robustness of HD map construction methods under multi-modal corrupted scenarios, we introduce two evaluation metrics.

Resilience Score (RS) We define RS as the relative robustness indicator for measuring how much accuracy a model can retain when evaluated on the corruption sets, which are calculated as follows:

$$\mathrm{RS}_i = \frac{\sum_{l=1}^{3} \mathrm{Acc}_{i,l}}{3 \times \mathrm{Acc}^{\mathrm{clean}}}, \quad \mathrm{mRS} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{RS}_i, \tag{1}$$

where $\mathrm{Acc}_{i,l}$ denotes the task-specific accuracy scores, with NDS (NuScenes Detection Score) for 3D object detection and mAP (mean Average Precision) for HD map construction, on corruption type i at severity level l. N is the total number of corruption types, and $\mathrm{Acc}^{\mathrm{clean}}$ denotes the accuracy score on the "clean" evaluation set. mRS (mean Resilience Score) represents the average score, providing an overall measure of the model's robustness across all types of corruption.

Relative Resilience Score (RRS) We define RRS as the critical metric for comparing the relative robustness of candidate models with the baseline model and mRRS as an overall metric to indicate the relative resilience score. The RRS and mRRS scores are calculated as follows:

$$\mathrm{RRS}_i = \frac{\sum_{l=1}^{3} \mathrm{Acc}_{i,l}}{\sum_{l=1}^{3} \mathrm{Acc}_{i,l}^{\mathrm{base}}} - 1, \quad \mathrm{mRRS} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{RRS}_i, \tag{2}$$

where $\mathrm{Acc}_{i,l}^{\mathrm{base}}$ denotes the accuracy of the baseline model.

IV. Method

Preliminaries For clarity, we first introduce the notation and definitions used throughout this paper. Our goal is to design a robust multi-modal HD map construction framework that integrates data augmentation, a novel multi-modal fusion module, and effective training strategies to significantly enhance the robustness of multi-modal fusion methods, as illustrated in Fig. 2. Formally, let $\chi = \{\mathit{Camera}, \mathit{LiDAR}\}$ represent the set of inputs, where $\mathit{Camera} \in \mathbb{R}^{B\times N^{cam}\times H^{cam}\times W^{cam}\times 3}$ denotes multi-view RGB camera images in perspective view (with B, N^{cam}, H^{cam}, and W^{cam} representing batch size, number of cameras, image height, and image width, respectively), and $\mathit{LiDAR} \in \mathbb{R}^{B\times P\times 5}$ represents the LiDAR point cloud (with P points, each containing 3D coordinates, reflectivity, and beam index). The detailed architectural designs are described in the following sections.

Data Augmentation To enhance robustness against sensor corruptions, we employ data augmentation strategies for both camera and LiDAR inputs. For camera data, we utilize GridMask [30], which randomly drops image information by applying a grid mask of the same size as the image, with binary values (0 or 1). For LiDAR data, we apply a dropout strategy [31] that randomly removes points from the point cloud to simulate sensor noise and improve model resilience.

After augmentation, we process the data as follows: For Camera Data, we utilize ResNet50 [32] as the backbone to extract multi-view features and apply GKT [33] as the 2D-to-BEV transformation module, converting these features into Bird's-Eye View (BEV) space. This results in BEV features $F_{Camera}^{BEV} \in \mathbb{R}^{B\times H\times W\times C}$, where H, W, and C denote height, width, and number of channels, respectively. For LiDAR Data, we follow the SECOND method [34] for voxelization and sparse LiDAR encoding. The resulting LiDAR features are projected into BEV space using a

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON NUSCENES VAL SET. L" AND C" REPRESENT LIDAR AND CAMERA, RESPECTIVELY. EFFI-B0", R50", PP", AND SEC" ARE SHORT FOR EFFICIENTNET-B0, RESNET50, POINTPILLARS AND SECOND, RESPECTIVELY. NOTE THAT ROBOMAP (MAPMODEL) MEANS OUR METHOD IS INTEGRATED INTO AN EXISTING MAPMODEL. BEST VIEWED IN COLOR.

Method	Venue	Modality	BEV Encoder	Backbone	Epoch	$\text{AP}_{ped.}$	$\text{AP}_{div.}$	$\text{AP}_{bou.}$	mAP ↑
HDMapNet[4]	ICRA'22	C	NVT	Effi-B0	30	14.4	21.7	33.0	23.0
VectorMapNet [5]	ICML'23	C	IPM	R50	110	36.1	47.3	39.3	40.9
PivotNet [21]	ICCV'23	C	PersFormer	R50	30	53.8	58.8	59.6	57.4
BeMapNet [22]	CVPR'23	C	IPM-PE	R50	30	57.7	62.3	59.4	59.8
MapVR [23]	NeurIPS'24	C	GKT	R50	24	47.7	54.4	51.4	51.2
MapTRv2 [24]	IJCV'24	C	BEVPoolv2	R50	24	59.8	62.4	62.4	61.5
StreamMapNet [25]	WACV'24	C	BEVFormer	R50	30	61.7	66.3	62.1	63.4
MapTR [6]	ICLR'23	C	GKT	R50	24	46.3	51.5	53.1	50.3
HIMap [26]	CVPR'24	C	BEVFormer	R50	24	62.2	66.5	67.9	65.5
VectorMapNet [5]	ICML'23	L	-	PP	110	25.7	37.6	38.6	34.0
MapTRv2 [24]	IJCV'24	L	-	Sec	24	56.6	58.1	69.8	61.5
MapTR [6]	ICLR'23	L	-	Sec	24	48.5	53.7	64.7	55.6
HIMap [26]	CVPR'24	L	-	Sec	24	54.8	64.7	73.5	64.3
HDMapNet [4]	ICRA'22	C & L	NVT	Effi-B0 & PP	30	16.3	29.6	46.7	31.0
VectorMapNet [5]	ICML'23	C & L	IPM	R50 & PP	110+ft	48.2	60.1	53.0	53.7
MBFusion [27]	ICRA'24	C & L	GKT	R50 & Sec	24	61.6	64.4	72.5	66.1
GeMap [28]	ECCV'24	C & L	GKT	R50 & Sec	24	66.3	62.2	71.1	66.5
MapTRv2 [24]	IJCV'24	C & L	BEVPoolv2	R50 & Sec	24	65.6	66.5	74.8	69.0
Mgmap [29]	CVPR'24	C & L	GKT	R50 & Sec	24	67.7	71.1	76.2	71.7
MapTR [6]	ICLR'23	C & L	GKT	R50 & Sec	24	55.9	62.3	69.3	62.5
HIMap [26]	CVPR'24	C & L	BEVFormer	R50 & Sec	24	71.0	72.4	79.4	74.3
RoboMap (MapTR)	-	**C & L**	**GKT**	**R50 & Sec**	**24**	**67.8**	**70.4**	**76.4**	**71.5**
RoboMap (HIMap)	-	**C & L**	**BEVFormer**	**R50 & Sec**	**24**	**74.6**	**74.5**	**82.0**	**77.0**

flattening operation as described in [35], yielding a unified LiDAR BEV representation $F_{LiDAR}^{BEV} \in \mathbb{R}^{B\times H\times W\times C}$.

Cross-modal Interaction Transform Existing methods convert sensory features into a shared BEV representation and fuse them to create multi-modal BEV features. However, LiDAR and camera features remain semantically misaligned due to modality gaps. To address this, we propose a Cross-Modal Interaction Transformer (CIT) module utilizing self-attention to enrich one modality with insights from another.

First, we start with the BEV features from both the camera ($F_{Camera}^{BEV} \in \mathbb{R}^{B\times H\times W\times C}$) and LIDAR ($F_{LiDAR}^{BEV} \in \mathbb{R}^{B\times H\times W\times C}$) sensors. The BEV tokens $\mathbf{T}_{\text{Camera}}^{\text{BEV}} \in \mathbb{R}^{HW\times C}$ and $\mathbf{T}_{\text{LiDAR}}^{\text{BEV}} \in \mathbb{R}^{HW\times C}$ are obtained by flattening each BEV feature and permuting the order of the matrices. Next, we concatenate the tokens of each modality and add a learnable positional embedding, which is a trainable parameter of dimension $2HW \times C$, to create the input BEV tokens $\mathbf{T}^{\text{in}} \in \mathbb{R}^{2HW\times C}$ for the Transformer. This positional embedding allows the model to distinguish spatial information between different tokens during training. Third, the input tokens $\mathbf{T}^{\text{in}}$ undergo linear projections to compute a set of queries, keys, and values ($\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$). Fourth, the self-attention layer computes the attention weights using scaled the dot product between $\mathbf{Q}$ and $\mathbf{K}$, and then multiplies these weights by the values to produce the refined output,

$$\mathbf{Z} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_{\text{k}}}}\right)\mathbf{V}, \quad (3)$$

where $\frac{1}{\sqrt{D_{\text{k}}}}$ is a scaling factor. To capture complex relationships across various representation subspaces and positions, we adopt the multi-head attention mechanism,

$$\hat{\mathbf{Z}} = \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\mathbf{Z}_1, \cdots, \mathbf{Z}_h)\mathbf{W}^{\text{O}}. \quad (4)$$

The subscript h denotes the number of head, and $\mathbf{W}^{\text{O}}$ denotes the projected matrix of $\text{Concat}(\mathbf{Z}_1, \cdots, \mathbf{Z}_h)$. Finally, the transformer uses a non-linear transformation to calculate the output features, $\mathbf{T}^{\text{out}}$ which are of the same shape as the input features $\mathbf{T}^{\text{in}}$,

$$\mathbf{T}^{\text{out}} = \text{MLP}(\hat{\mathbf{Z}}) + \mathbf{T}^{\text{in}}. \quad (5)$$

The output $\mathbf{T}^{\text{out}}$ are converted into $\hat{\mathbf{F}}_{\text{Camera}}^{\text{BEV}}$ and $\hat{\mathbf{F}}_{\text{LiDAR}}^{\text{BEV}}$ for further feature fusion. We utilize the Dynamic Fusion module to aggregate the multi-modal BEV feature inputs, $\hat{\mathbf{F}}_{\text{Camera}}^{\text{BEV}}$ and $\hat{\mathbf{F}}_{\text{LiDAR}}^{\text{BEV}}$, resulting in the aggregated features $\mathbf{F}_{\text{fused}}$. The output fused feature $\mathbf{F}_{fused}$ will be used for HD Map construction task, with the decoder and prediction heads.

Modality Dropout Training Strategy To simulate real-world sensor failures during training, we employ a Modality Dropout strategy, where the BEV features of either the camera or LiDAR ($\hat{\mathbf{F}}_{\text{Camera}}^{\text{BEV}}$ or $\hat{\mathbf{F}}_{\text{LiDAR}}^{\text{BEV}}$) are randomly dropped with a probability p_{md}. When a modality is dropped, p_{L} denotes the probability of retaining the LiDAR input, while $p_{\text{C}} = 1 - p_{\text{L}}$ represents the probability of retaining the camera input. Thus, the overall probability distribution is as follows: the probability of retaining both sensors is $1-p_{\text{md}}$, the probability of retaining only LiDAR is $p_{\text{md}} \cdot p_{\text{L}}$, and the probability of retaining only the camera is $p_{\text{md}} \cdot (1 - p_{\text{L}})$. This strategy enhances the model's robustness to partial sensor failures by randomly dropping modalities, enabling it to better adapt to real-world scenarios where sensor malfunctions may occur.

TABLE II

THE SCORES RS_c AND mRS FOR THE ORIGINAL MAPTR [6] MODEL AND ITS VARIANTS. RS_c USING MAP AS METRIC.

Model	Motion Blur	Temporal Mis.	Spatial Mis.	Fog	Snow	Camera Crash	Frame Lost	Cross Sensor	Cross Talk	Camera Crash, Cross Sensor	Camera Crash, Cross Talk	Frame Lost, Cross Sensor	Frame Lost, Cross Talk	mRS↑
MapTR (Baseline)	70.00	76.94	69.05	67.94	19.55	78.69	74.75	98.47	84.99	77.40	58.14	73.50	54.27	69.51
MapTR (Baseline) + Fusion Module	80.03	75.28	68.48	68.26	23.67	70.77	64.15	96.34	87.93	69.52	56.57	62.94	51.03	67.31
MapTR (Baseline) + Data Augmentation	71.70	75.04	68.16	66.76	24.66	79.29	76.88	96.63	90.31	77.81	66.92	75.42	63.86	71.80
MapTR (Baseline) + Dropout Training	72.11	73.52	57.10	63.19	20.59	82.23	80.55	94.09	81.13	80.34	64.06	78.67	62.00	69.97
RoboMap (MapTR)	**89.88**	**73.48**	**63.02**	**69.17**	**23.78**	**93.49**	**92.86**	**95.90**	**86.83**	**91.85**	**78.95**	**91.07**	**77.56**	**79.06**

TABLE III

THE SCORES RS_c AND mRS FOR THE ORIGINAL HIMAP [26] MODEL AND ITS VARIANTS. RS_c USING MAP AS METRIC.

Model	Motion Blur	Temporal Mis.	Spatial Mis.	Fog	Snow	Camera Crash	Frame Lost	Cross Sensor	Cross Talk	Camera Crash, Cross Sensor	Camera Crash, Cross Talk	Frame Lost, Cross Sensor	Frame Lost, Cross Talk	mRS↑
HIMap (Baseline)	83.77	74.93	77.31	75.56	23.79	63.41	58.59	97.84	94.28	61.91	54.15	57.19	57.19	67.69
HIMap (Baseline) + Fusion Module	83.69	72.92	79.06	75.17	25.59	68.58	63.92	97.69	94.86	67.15	60.57	62.56	56.16	69.84
HIMap (Baseline) + Data Augmentation	84.35	73.30	79.21	75.44	25.50	66.27	61.11	97.52	94.93	64.81	58.63	59.64	53.63	68.80
HIMap (Baseline) + Dropout Training	84.21	72.09	71.73	70.93	27.13	87.41	85.15	96.49	91.50	85.69	76.56	83.39	72.92	77.32
RoboMap (HIMap)	**90.34**	**72.54**	**72.32**	**76.69**	**30.44**	**93.17**	**92.58**	**97.21**	**93.24**	**91.68**	**83.26**	**90.95**	**81.43**	**81.99**

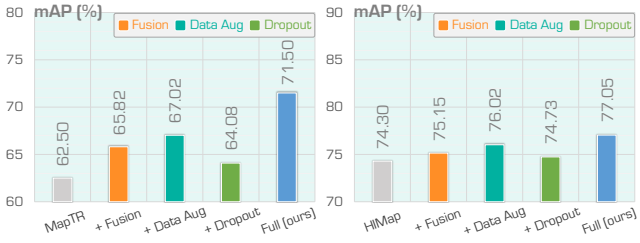

Fig. 3. Analyze the impact of different modules on the HD map construction task using clean data.

V. EXPERIMENTS AND ANALYSIS

A. Experimental Settings

Dataset The nuScenes dataset [20] consists of 1,000 sequences collected by autonomous vehicles. Each sample is annotated at 2Hz and includes six camera images capturing the 360° horizontal field of view of the ego-vehicle. Following the methodologies in [6], [26], we focus on three key map elements: pedestrian crossings, lane dividers, and road boundaries, to ensure a fair evaluation.

Evaluation Metrics For clean data, we adopt metrics consistent with prior HD map studies [6], [26]. Average Precision (AP) measures the quality of map construction, while **Chamfer Distance** ($D_{\textbf{Chamfer}}$) quantifies the alignment between predictions and ground truth. To assess model robustness, we introduce the Resilience Score (RS) and Relative Resilience Score (RRS), which evaluate the model's performance under data corruption or sensor noise, ensuring reliability in real-world scenarios.

Implementation Details Our RoboMap framework is trained on four NVIDIA RTX A6000 GPUs. We retrain two state-of-the-art baseline models, MapTR [6] and HIMap [26], using their official configurations from open-source repositories. All experiments employ the AdamW optimizer with a learning rate of 4.2×10^{-4}. Notably, RoboMap's core components—data augmentation, multi-modal fusion module, and training strategies—are designed as simple yet effective plug-and-play techniques, making them compatible with existing camera-LiDAR fusion pipelines for HD map construction.

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

With the same settings and data partition, we compare the proposed RoboMap model with several state-of-the-art methods, including HDMapNet [4], VectorMapNet [5], MBFusion [27], GeMap [28], MgMap [29], MapTR [6], MapTRv2 [24], and HIMap [26]. The overall performance of RoboMap and all baselines on the nuScenes dataset is summarized in Tab. I.

The experimental results highlight several key observations: multi-modal approaches consistently outperform single-modal methods, demonstrating the importance of leveraging complementary information from both camera and LiDAR sensors for HD map construction. As shown in Tab. I, RoboMap achieves significant improvements over the original models, with RoboMap (MapTR) surpassing the original camera-LiDAR fusion MapTR model by 9 mAP on the nuScenes dataset and RoboMap (HIMap) outperforming the previous state-of-the-art HIMap fusion model by 2.7 mAP, setting a new benchmark for vectorized map reconstruction. The superior performance of RoboMap can be attributed to its three core components—data augmentation, a multi-modal fusion module, and advanced training strategies—which collectively enhance robustness and accuracy. In summary, RoboMap demonstrates substantial superiority over existing multi-modal methods, highlighting its effectiveness in HD map construction tasks.

C. Ablation Studies

To systematically evaluate the effectiveness of each component in our proposed RoboMap, we conduct ablation studies by incrementally adding individual strategies to the baseline model and present the results in Fig. 3. Specifically, we design the following ablation models: (1) **RoboMap (w/o Fusion)**, which integrates a cross-modal interaction transformation fusion module into the original baseline model; (2) **RoboMap (w/ Data Augmentation)**, which incorporates image and LiDAR data augmentation strategies into the original baseline model; (3) **RoboMap (w/ Dropout)**, which applies the Modality Dropout Training Strategy to the original baseline model; and (4) **RoboMap (full)**, which combines all three key components—data augmentation, a multi-modal fusion

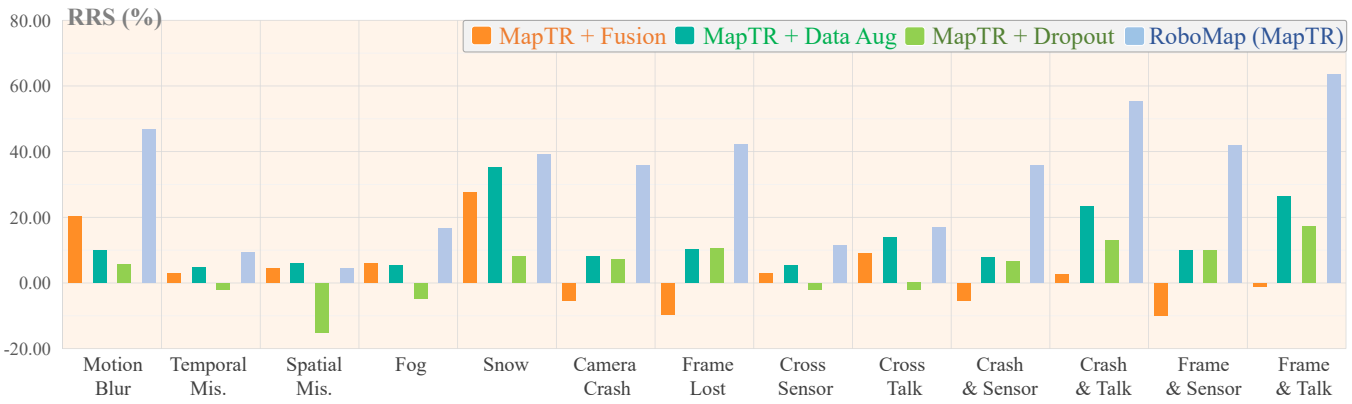

Fig. 4. Relative robustness visualization. Relative Resilience Score (RRS) computed with mAP using original MapTR [6] as baseline.

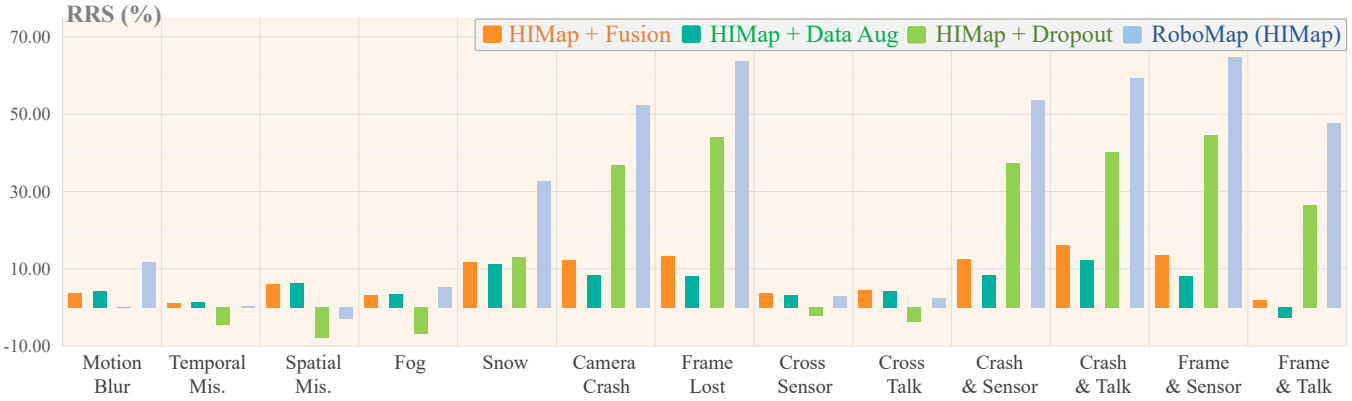

Fig. 5. Relative robustness visualization. Relative Resilience Score (RRS) computed with mAP using original HIMap [26] as baseline.

module, and training strategies—into the baseline model.

The ablation results demonstrate that each component significantly enhances the baseline model's performance. Specifically, RoboMap (w/ Fusion), RoboMap (w/ Data Augmentation), and RoboMap (w/ Dropout) outperform the baseline MapTR model on the nuScenes dataset, achieving gains of 3.3, 4.5, and 1.6 mAP, respectively. Similarly, these variants surpass the state-of-the-art HIMap model, with improvements of 0.85, 1.7, and 0.4 mAP, respectively. These extensive experimental results validate the effectiveness of each strategy in improving model performance, highlighting the robustness and versatility of RoboMap.

D. Robustness of multi-sensor corruptions

To explore strategies that enhance robustness, such as data augmentation, multi-modal fusion, and modality dropout training, we evaluated the popular MapTR [6] and the state-of-the-art HIMap [26] models. Tab. II and Tab. III present their Resilience Scores, while Fig. 4 and Fig. 5 illustrate their Relative Resilience Scores. Our analysis reveals two key insights. First, while camera-LiDAR fusion methods show promising performance by integrating multi-modal data, many approaches assume complete sensor availability, leading to low robustness when sensors are corrupted or missing. Second, although individual strategies do not consistently improve robustness across all multi-sensor corruption scenarios, combining them significantly enhances model resilience. Specifically, our approach improves the mRS metric by 9.55 and 14.3 compared to the original MapTR and HIMap models, respectively, demonstrating the effectiveness of these strategies in boosting robustness.

The experimental results emphasize the need to address sensor vulnerabilities in multi-modal systems. While camera-LiDAR fusion performs well under ideal conditions, its dependence on complete sensor data makes it prone to failure in real-world scenarios with incomplete or corrupted data. By incorporating data augmentation, multi-modal fusion, and modality dropout training, we significantly improve robustness. These strategies enhance the resilience of both MapTR and HIMap models and offer a framework for building more robust multi-modal systems. The findings highlight the potential of targeted enhancements to tackle real-world challenges in sensor-based applications.

VI. Conclusion

In this paper, we improve the robustness of HD map construction methods, essential for autonomous driving systems. We propose a comprehensive framework integrating data augmentation, a multi-modal fusion module, and innovative modality dropout training strategies. Experimental results demonstrate our method significantly enhances robustness on a dataset with 13 types of sensor corruption. Additionally, our approach achieves state-of-the-art performance on the clean dataset. Overall, our model offers valuable insights for developing more reliable HD map techniques, contributing to safer and more effective autonomous driving technologies.

REFERENCES

[1] X. Hao, R. Li, H. Zhang, D. Li, R. Yin, S. Jung, S.-I. Park, B. Yoo, H. Zhao, and J. Zhang, "Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation," in *European Conference on Computer Vision*, 2024, pp. 166–183.

[2] S. Wang, F. Jia, W. Mao, Y. Liu, Y. Zhao, Z. Chen, T. Wang, C. Zhang, X. Zhang, and F. Zhao, "Stream query denoising for vectorized hd-map construction," in *European Conference on Computer Vision*, 2024, pp. 203–220.

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

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

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

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

[7] X. Hao, Y. Diao, M. Wei, Y. Yang, P. Hao, R. Yin, H. Zhang, W. Li, S. Zhao, and Y. Liu, "Mapfusion: A novel bev feature fusion network for multi-modal map construction," *Information Fusion*, 2025.

[8] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," 2022, pp. 10 421–10 434.

[9] X. Wang, C. Fu, Z. Li, Y. Lai, and J. He, "Deepfusionmot: A 3d multi-object tracking framework based on camera-lidar fusion with deep association," *IEEE Robotics and Automation Letters*, pp. 8260–8267, 2022.

[10] M. Seong, J. Kim, G. Bang, H. Jeong, and J. W. Choi, "Mr-occ: Efficient camera-lidar 3d semantic occupancy prediction using hierarchical multi-resolution voxel representation," *arXiv preprint arXiv:2412.20480*, 2024.

[11] X. Hao, Y. Yang, H. Zhang, M. Wei, Y. Zhou, H. Zhao, and J. Zhang, "Team samsung-ral: Technical report for 2024 robodrive challenge-robust map segmentation track," *arXiv preprint arXiv:2405.10567*, 2024.

[12] L. Kong, S. Xie, H. Hu, Y. Niu, W. T. Ooi, B. R. Cottereau, L. X. Ng, Y. Ma, W. Zhang, L. Pan *et al.*, "The robodrive challenge: Drive anytime anywhere in any condition," *arXiv preprint arXiv:2405.08816*, 2024.

[13] X. Hao, Y. Yang, H. Zhang, M. Wei, Y. Zhou, H. Zhao, and J. Zhang, "Using temporal information and mixing-based data augmentations for robust hd map construction."

[14] S. Xie, L. Kong, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robobev: Towards robust bird's eye view perception under corruptions," *arXiv preprint arXiv:2304.06719*, 2023.

[15] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[16] L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 994–20 006.

[17] Z. Zhu, Y. Zhang, H. Chen, Y. Dong, S. Zhao, W. Ding, J. Zhong, and S. Zheng, "Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 600–21 610.

[18] X. Hao, M. Wei, Y. Yang, H. Zhao, H. Zhang, Y. Zhou, Q. Wang, W. Li, L. Kong, and J. Zhang, "Is your hd map constructor reliable under sensor corruptions?" *Conference on Neural Information Processing Systems*, 2024.

[19] X. Hao, G. Liu, Y. Zhao, Y. Ji, M. Wei, H. Zhao, L. Kong, R. Yin, and Y. Liu, "Msc-bench: Benchmarking and analyzing multi-sensor corruption for driving perception," *arXiv preprint arXiv:2501.01037*, 2025.

[20] H. Caesar, V. Bankiti *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 618–11 628.

[21] W. Ding, L. Qiao, X. Qiu, and C. 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.

[22] 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*, 2023, pp. 13 218–13 228.

[23] Z. Zhang, Y. Zhang, X. Ding, F. Jin, and X. Yue, "Online vectorized hd map construction using geometry," in *European Conference on Computer Vision*, 2024, pp. 73–90.

[24] 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," *International Journal of Computer Vision*, pp. 1–23, 2024.

[25] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. 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.

[26] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 396–15 406.

[27] X. Hao, H. Zhang, Y. Yang, Y. Zhou, S. Jung, S.-I. Park, and B. Yoo, "Mbfusion: A new multi-modal bev feature fusion method for hd map construction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 922–15 928.

[28] Z. Zhang, Y. Zhang, X. Ding, F. Jin, and X. Yue, "Online vectorized hd map construction using geometry," in *European Conference on Computer Vision*, 2024, pp. 73–90.

[29] X. Liu, S. Wang, W. Li, R. Yang, J. Chen, and J. 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. 14 812–14 821.

[30] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[31] J. Choi, Y. Song, and N. Kwak, "Part-aware data augmentation for 3d object detection in point cloud," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 3391–3397.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[34] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, p. 3337, 2018.

[35] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE international conference on robotics and automation (ICRA)*, 2023, pp. 2774–2781.