

ROSE: Roadside Oversight-Guided Scenario Enhancement with Self-Supervised Coupling for multi-modal Perception

Guoyu Zhang¹, Peng Hang^{1,✉}, Xin Xia², Jian Sun¹

 Cite this article: <https://doi.org/10.26599/COMMTR.2026.9640052>

ABSTRACT: With the rapid advancement of autonomous driving and Intelligent Transportation Systems (ITS), roadside perception—an essential component of vehicle-to-everything (V2X) communication—has become a critical foundation for large-scale traffic monitoring and data-driven safety decisions. However, under adverse environmental conditions such as rain, snow, fog, and nighttime, current roadside perception systems often face limited training data, high annotation costs, and significant performance degradation due to poor model robustness. To tackle these challenges without relying on additional labeled data, this paper introduces ROSE (Roadside Oversight-guided Scenario Enhancement), a unified closed-loop framework designed to enhance the resilience and adaptability of roadside multi-modal perception systems. ROSE integrates three key components: (1) RISA, a first-order physics-guided cross-modal augmentation module that generates physically plausible and semantically aligned adverse-weather samples; (2) SSL Coupling, a cross-modal self-supervised learning network that facilitates robust feature alignment; and (3) a curriculum scheduling mechanism guided by vision-language models (VLM) to adaptively prioritize learning difficulty. Experimental results show that ROSE achieves competitive detection accuracy while providing improved robustness and cross-modal consistency across adverse weather scenarios, demonstrating strong generalization potential for roadside multi-modal perception under challenging environmental conditions. These findings suggest that ROSE offers a practical and effective pathway toward building resilient V2X cooperative perception systems.

KEYWORDS: Roadside Perception Systems, multi-modal Fusion, Self-Supervised Learning, Vision-Language Model, Adaptive

1 Introduction

The advancement of Intelligent Transportation Systems (ITS) has made roadside perception a critical component for ensuring the safety and efficiency of vehicle-to-everything (V2X) cooperation (Gao et al., 2025). By deploying a diverse array of sensors such as cameras and LiDAR on roadside infrastructure, these systems can effectively extend a vehicle's perception range, eliminate blind spots, and provide comprehensive coverage of the traffic environment (Asif-Ur-Rahman et al., 2019; Li et al., 2024; Fang et al., 2026). This capability is particularly vital under adverse weather conditions—such as rain, snow, fog, and nighttime—where onboard sensors typically suffer from limited detection range (usually less than 150 meters) and are highly susceptible to occlusion, weather interference, and lighting variations, resulting in substantial degradation of perception performance (Kim et al., 2023; Zhang et al., 2023). Such conditions introduce disruptive noise into sensor data—e.g., occlusions, scattering, and artifacts—that can cause missed or false detections, posing a significant bottleneck for intelligent driving systems (Aloufi et al., 2024; Bijelic et al., 2020; Lu et al., 2014; Musat et al., 2021).

multi-modal perception fusion is widely regarded as an effective approach to improving robustness under adverse weather conditions (Li et al., 2025). Different sensors offer complementary information and performance characteristics: cameras provide rich

texture and semantic details but are sensitive to lighting variations (Zhang et al., 2025); millimeter-wave radar offers strong penetration but suffers from low resolution; LiDAR delivers high-precision 3D measurements with minimal sensitivity to illumination, yet is expensive and still vulnerable to extreme weather disturbances (Sezgin et al., 2023; Zang et al., 2019). Integrating these modalities enables the system to leverage their respective strengths and mitigate the limitations of any single sensor modality in extreme scenarios, making multi-modal fusion a crucial research direction toward all weather perception systems (Li et al., 2022; Liang et al., 2022; Vora et al., 2020; Yin et al., 2024).

Despite these advantages, achieving robust multi-modal perception in adverse weather remains challenging. A major barrier lies in the scarcity of large-scale annotated datasets that include extreme weather conditions (El-Shair et al., 2024; Sheeny et al., 2020). Due to high data collection and labeling costs, existing public datasets severely under-represent complex weather scenarios and often suffer from annotation noise. Moreover, small-scale datasets focused on specific weather types tend to exhibit domain distribution shifts, making them unsuitable for training models with strong generalization capability. As a result, many models that perform well under clear weather show significant performance drops in rain or snow (Choi et al., 2025; Duy-Tho Le, 2025; Shi et al., 2025).

¹ College of Transportation, Tongji University, Shanghai 201804, China. ² Department of Mechanical Engineering, University of Michigan Dearborn, Dearborn 48188, USA.

✉ Corresponding author. E-mail: hangpeng@tongji.edu.cn

Received: January 6, 2026; Revised: March 12, 2026; Accepted: September 7, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

This highlights the need for perception enhancement methods that do not rely heavily on large-scale labeled data and can generalize across diverse weather conditions.

Recent efforts to address these challenges have focused on three key directions: optimizing multi-modal fusion strategies, enhancing perception robustness, and constructing alternative training data. In the fusion domain, studies have explored deep integration of heterogeneous data sources—including visible light, LiDAR, millimeter-wave radar, and infrared—especially under low-visibility conditions, to improve object detection and recognition (Cho et al., 2025; Massoud and Laganieri, 2024; Qian et al., 2025; Sakaridis et al., 2021; Vinoth and Sasikumar, 2024). To combat severe weather interference, researchers have adopted synthetic point cloud or image generation to train perception models, boosting their robustness under conditions such as heavy snow or dense fog (M. Hahner, 2021; S. Teufel, 2023; Zhao et al., 2025). Additionally, noise modeling and denoising techniques for corrupted LiDAR point clouds have been shown to improve data quality and subsequent 3D detection performance.

Furthermore, data augmentation and domain adaptation techniques have been widely employed to simulate extreme weather scenarios (You et al., 2025). Cross-modal augmentation and style transfer methods can transform clear-weather data into versions simulating rain, fog, or snow, thus expanding training diversity and reducing performance degradation due to domain shifts (Kumar and Muhammad, 2023; Wang et al., 2021). Scene-aware deweathering and reconstruction methods, incorporating semantic and depth information, also enhance object visibility and improve the consistency of perception outputs under adverse conditions (Chen et al., 2023; Wang et al., 2023; Vilsan. A et al., 2025).

The rise of self-supervised learning has further emphasized the value of unlabeled data in perception model training. By modeling cross-modal consistency—e.g., geometric alignment and feature matching between cameras and LiDAR—it is possible to extract effective supervisory signals without manual labels (Afham et al., 2022). Cross-modal contrastive pretraining, followed by downstream fine-tuning, has demonstrated superior performance in various 3D perception tasks, underscoring the feasibility of building robust perception systems based on large-scale, unlabeled multi-modal data (Afham et al., 2022; Xu et al., 2021; Yang et al., 2021).

Recently, Large Language Models (LLMs) and Vision-Language Models (VLMs) have broadened the role of foundation models in intelligent transportation systems by linking visual perception with semantic reasoning and decision support. In autonomous driving, VLMs have been investigated for tasks such as scene description, traffic-context interpretation, risk assessment, and hierarchical planning, with particular advantages in complex or long-tail scenarios (Liu et al., 2025). Beyond autonomous driving, LLMs have also been applied to a wider range of ITS tasks, including traffic prediction and control, safety analysis, simulation, and knowledge-assisted decision optimization (Lu et al., 2025; Liu et al., 2023). Collectively, these efforts indicate that language-enhanced foundation models can complement conventional perception networks with richer semantic knowledge and reasoning capability. Nevertheless, current VLM/LLM-based research is largely centered on scene-level understanding, reasoning, and planning. Their role in improving low-level multi-modal perception, particularly when sensor quality deteriorates under adverse weather, has received much less attention. In this work, we therefore explore a different use of VLMs: instead of placing them directly in the online perception pipeline, we employ them as semantic schedulers that adapt the training process according to scene difficulty, modality degradation, and the observed performance of multi-modal fusion.

Overall, previous research has approached adverse-weather perception from several directions, including multi-modal fusion, data

augmentation, self-supervised learning, and more recently, language-enhanced foundation models. Each line of work addresses a different part of the problem. Multi-modal fusion compensates for the limitations of individual sensors, while data augmentation and self-supervised learning reduce the dependence on large amounts of labeled adverse-weather data. VLM- and LLM-based approaches further introduce semantic reasoning and adaptive decision capabilities. However, these techniques are commonly developed as separate solutions for individual stages of the perception pipeline. Few existing frameworks jointly consider physically consistent degradation modeling, collaborative learning across modalities, and training-policy adaptation. Consequently, the combined challenges of limited adverse-weather data, heterogeneous sensor degradation, and changing training difficulty are still not fully addressed within a single framework.

To bridge this gap, we propose ROSE (Roadside Oversight-guided Scenario Enhancement), a robustness-oriented framework for roadside multi-modal perception. ROSE combines physically plausible data augmentation, cross-modal self-supervised learning, and curriculum-based adaptive scheduling within one training framework. Its contribution does not rely on introducing an entirely new learning primitive at every stage. Instead, ROSE emphasizes the coordinated interaction among physics-guided augmentation, cross-modal representation learning, and feedback-driven curriculum adaptation for adverse-weather roadside perception. The three components communicate through explicit data and policy interfaces. RISA defines a controllable domain of sensor degradation; SSL Coupling turns clean and augmented data pairs into learning signals that encourage robust cross-modal representations; and an outer-loop scheduler uses training feedback to modify the sampling and augmentation policy for subsequent training rounds. In this way, ROSE jointly targets three practical difficulties—insufficient adverse-weather data, inconsistent degradation across sensing modalities, and the limitations of a fixed training curriculum. The main contributions are summarized as follows:

- 1) Physically plausible cross-modal augmentation module (RISA): RISA generates adverse-weather conditions, including rain, snow, fog, and nighttime, for both image and point-cloud data using physical models and environmental priors. Instead of independently perturbing each sensor stream, the same environmental conditions are reflected across modalities so that their spatial correspondence is preserved after augmentation. Compared with conventional augmentation strategies that are typically designed for a single modality or rely mainly on appearance-level perturbations, RISA expands the training distribution through physically grounded cross-modal transformations. This provides the detector with more diverse degraded observations while maintaining the consistency required for multi-modal fusion.
- 2) Cross-modal self-supervised coupled learning strategy: We develop a teacher-student self-supervised learning scheme that combines cross-modal contrastive learning with feature- and prediction-level consistency constraints. The objective is to exploit complementary information between images and point clouds while reducing the dependence on additional manual annotations. By encouraging consistent representations and predictions across modalities and between clean and degraded observations, the proposed strategy helps the model retain useful cross-modal information when part of the sensory input is corrupted. This improves robustness and generalization under challenging environmental conditions.
- 3) Constrained VLM-assisted closed-loop curriculum scheduling: A VLM-assisted curriculum scheduler is introduced to link model feedback from the current training stage with augmentation and sampling decisions in subsequent stages. The VLM does not

directly update detector parameters. Instead, it operates in an outer loop, where it analyzes failed samples, augmentation configurations, and historical fusion behavior, and then produces structured curriculum recommendations. These recommendations are subsequently translated into bounded updates of bucket-level sampling policies subject to physical-validity and semantic-consistency constraints. Such a constrained update mechanism allows training difficulty to evolve with model performance while avoiding abrupt or semantically unreasonable policy changes. As a result, data generation, representation learning, and curriculum adaptation form a feedback-driven training loop rather than three independent procedures.

The remainder of this paper is organized as follows. Sections II and III describe the ROSE framework and its main components, including cross-modal augmentation, self-supervised coupled learning, and curriculum scheduling. Section IV introduces the experimental setup and evaluates the proposed framework under different adverse-weather conditions, followed by ablation studies examining the contribution of individual components. Finally, Section V concludes the paper and discusses directions for future work.

2 Framework

In intelligent transportation scenarios, roadside perception systems are frequently exposed to adverse conditions, including rain, snow, fog, abrupt illumination changes, and partial occlusions. Conventional

perception models are often trained with large amounts of annotated data under relatively fixed training schemes, which limits their ability to represent the diversity of these long-tail conditions (Wang et al., 2024a). To improve robustness in such settings, we develop ROSE as a closed-loop training framework that coordinates physics-guided augmentation, cross-modal self-supervised learning, and curriculum adaptation over successive training rounds.

Rather than treating data generation, representation learning, and training scheduling as separate stages, ROSE links them through a common feedback process. The framework starts from real roadside multi-modal observations and uses RISA to construct physically grounded paired samples under simulated adverse conditions. These clean and augmented data pairs are then used within a self-supervised coupling scheme to provide robustness-oriented learning signals across modalities. Training statistics collected during this process are subsequently passed to a VLM-based scheduler, which adjusts the sampling and augmentation strategy for the following round. As training proceeds, the system therefore evolves through an iterative chain of “samples → metadata → statistics → strategy,” allowing the training distribution to adapt to model behavior while preserving the physical and semantic consistency of the generated data.

Through this design, ROSE improves the robustness and generalization of multi-modal roadside perception without requiring additional manual annotations. An overview of the complete framework is shown in Figure 1.

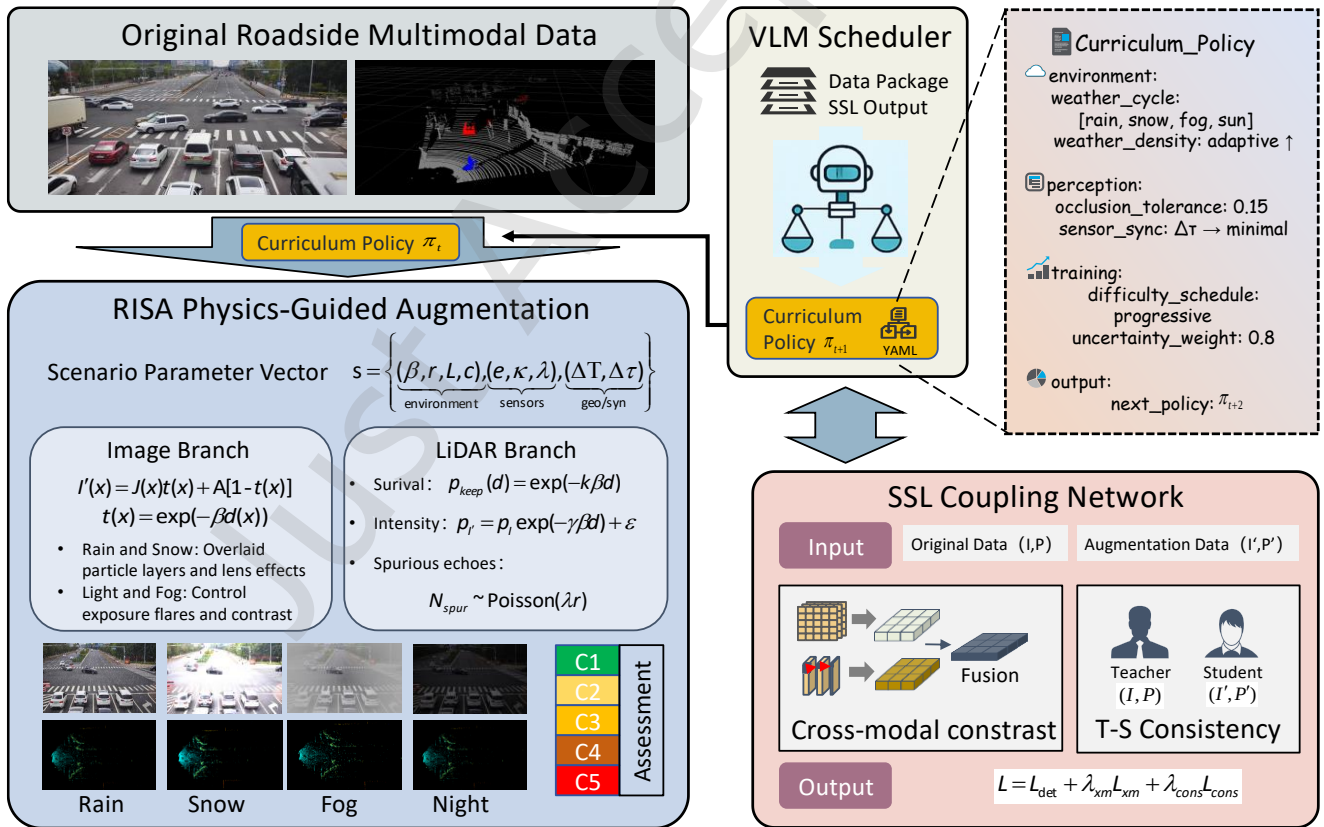


Fig. 1. Overall framework of ROSE. Overall architecture of the ROSE framework, integrating physics-guided augmentation (RISA), self-supervised learning (SSL), and curriculum-based training driven by the VLM scheduler. The process begins with original multi-modal data and progresses through augmentation, self-supervised learning, and dynamic curriculum scheduling for enhanced multi-modal perception.

At the data level, ROSE starts from synchronized Camera - LiDAR observations acquired by roadside sensing systems. Before augmentation, the two sensing streams are temporally aligned and spatially calibrated. Sample selection and degradation severity are then

controlled by an initial curriculum policy. RISA describes the environmental conditions, sensor characteristics, and cross-modal geometry through a common set of scene parameters. For the camera branch, rain, snow, fog, and low-light conditions are synthesized using

atmospheric scattering and optical imaging models. LiDAR corruption is generated separately, but under the same environmental setting, by introducing effects such as range-dependent point loss, intensity decay, and spurious returns. Because both modalities share the same scene-level conditions, the resulting degraded Camera - LiDAR pairs preserve their spatial correspondence while reflecting physically reasonable sensing deterioration.

During training, both the clean observations and their augmented versions are fed into a self-supervised coupled learning framework built on top of a standard multimodal detector. The framework introduces two complementary learning objectives. Cross-modal contrastive learning is used to maintain the correspondence between camera and LiDAR representations of the same object, even when one of the sensing streams is strongly degraded. This helps preserve both geometric and semantic consistency in the latent feature space. In parallel, a teacher - student consistency mechanism is employed: the teacher network processes clean inputs, whereas the student network receives degraded observations. Consistency constraints at the feature or prediction level encourage the student to retain stable responses under adverse conditions. Combined with the supervised detection objective, these losses allow the model to benefit from labeled as well as weakly labeled or unlabeled samples, while improving robustness to rain, fog, poor illumination, occlusion, and sparse LiDAR measurements.

Rather than sending raw optimization information to the VLM, ROSE periodically summarizes the current training state into structured metadata. These summaries contain global and scenario-level loss statistics, modality confidence, dominant error categories, failure indices, augmentation settings, and sample-difficulty information. The VLM scheduler interprets this information between training rounds and adjusts the curriculum according to the weaknesses observed in the current model. Depending on the failure pattern, it may change the sampling probability of particular degradation conditions, revise sample-retention rules, or alter the rate at which more difficult examples are introduced. Proposed parameter combinations that violate physical or semantic constraints are rejected, while degradation scenarios associated with persistent detection failures may be sampled more frequently. Once validated, the updated policy is translated into an explicit configuration for RISA and data selection in the following round. The training distribution can therefore adapt to the model's evolving failure modes instead of following a fixed curriculum throughout optimization.

To avoid unnecessary coupling among augmentation, model learning, and curriculum scheduling, ROSE adopts a unified sample-package interface. Each package contains the original and augmented Camera - LiDAR observations, aligned annotations, timestamps, calibration information, scene variables, difficulty labels, the curriculum version, and a random seed. Cached intermediate results may also be included when visualization or error analysis is required. The packages can be passed from augmentation to training through pipelined queues or zero-copy buffers, whereas the VLM only receives compact summaries of the training state rather than the underlying samples.

This design also makes the role of the VLM more explicit. It is not involved in backpropagation and never updates detector parameters directly. Instead, it acts outside the gradient-based optimization loop as a constrained scheduler operating over predefined curriculum buckets. Its function is therefore to provide bounded and valid curriculum adjustments, rather than to solve or guarantee convergence of the underlying non-convex detector optimization problem. Since the scheduler is invoked only at designated curriculum-update stages and reasons over aggregated training statistics, it does not lie on the per-iteration optimization path and remains only loosely coupled with detector training.

The use of a VLM still introduces uncertainty into curriculum

adaptation. For example, the model may incorrectly attribute a particular detection failure to the wrong sensing condition and consequently recommend an inappropriate sampling adjustment. ROSE reduces this risk by inserting deterministic validation steps before any suggested policy is executed. A physical-validity check first determines whether the proposed update is compatible with the predefined environmental and sensor-degradation models. A semantic-consistency check then evaluates whether the recommendation agrees with the observed training evidence, including changes in detection performance, cross-modal fusion inconsistency, and teacher - student disagreement. Only proposals satisfying these checks are accepted. Otherwise, the previous curriculum is preserved or replaced by a conservative fallback policy. Curriculum adaptation is thus driven jointly by VLM-based interpretation and explicit rule-based constraints rather than by unconstrained language-model decisions. Accepted updates are logged together with the corresponding physical and semantic evidence so that each scheduling decision can be traced during training.

Taken together, ROSE organizes physically grounded augmentation, cross-modal self-supervised learning, and adaptive curriculum scheduling into a single training pipeline built around real roadside Camera - LiDAR data. Training alternates among data augmentation, model optimization, evaluation, and curriculum revision. As the curriculum changes, the model is gradually exposed to a broader range of degradation conditions, while its cross-modal representations become less sensitive to deterioration in either sensing modality. Importantly, the augmentation and scheduling components are used only during training. At inference time, roadside 3D detection relies solely on the trained perception network, without introducing RISA or the VLM scheduler into the online processing path. This keeps deployment lightweight while retaining the robustness acquired through adverse-weather augmentation and adaptive training.

3 Methodology

The proposed ROSE framework combines supervised-guided scenario augmentation with self-supervised coupled training for robust roadside multimodal perception. It is built around three main components. First, the physically interpretable augmentation module, RISA, introduces adverse weather conditions, such as rain, fog, and snow, into camera images and LiDAR point clouds while maintaining temporal consistency across the two modalities. Second, a cross-modal self-supervised learning (SSL) module is incorporated into the multimodal 3D object detection network, where contrastive learning and teacher - student consistency are jointly used to improve model robustness. Third, a VLM-based scheduler adjusts the curriculum and augmentation settings during training according to heuristic rules and feedback from the model's training performance. The three components are described in detail in the following sections.

3.1. RISA Physics-Guided Augmentation

To improve robustness to long-tail roadside conditions such as rain, snow, and fog without introducing extra annotation effort, we develop RISA, a physics-guided cross-modal data augmentation module. A shared scene parameter vector is used to jointly control environmental, sensor-level, and geometric augmentations.

$$s = \{(\beta, r, L), (e, \kappa, \lambda), (\Delta T, \Delta \tau)\} \quad (1)$$

The unified scene parameter vector s in RISA encapsulates the coupled modeling of environment, sensor, and geometric synchronization factors that jointly drive physically plausible cross-modal augmentation. Environmental parameters characterize objective scene conditions such as visibility, precipitation intensity, and lighting, which jointly influence both image and LiDAR degradation. Specifically, the atmospheric extinction coefficient β , precipitation particle density r , and illumination level L serve as key variables controlling the severity

and type of environmental perturbations applied across modalities.

Sensor parameters describe the intrinsic characteristics and noise profiles of the camera and LiDAR hardware. The electronic noise level of the camera is denoted as e , while LiDAR specific degradation is governed by the range attenuation scaling factor κ , which modulates the intensity decay with respect to distance, and the pseudo-echo baseline rate λ , which affects the density and distribution of spurious points caused by back scattering in harsh weather conditions.

Geometric synchronization parameters model temporal and spatial misalignments between camera and LiDAR acquisitions, which are critical for simulating realistic data inconsistencies and training robust cross-modal alignment. The intersensor temporal offset ΔT represents the global timestamp deviation between the camera's exposure center and the LiDAR's time reference, while the scanning phase bias Δt reflects the angular misalignment between the two sensors' acquisition sequences. By adjusting these parameters in a controlled manner, RISA enables the simulation of real-world asynchronous distortions and facilitates training under realistic multi-sensor timing uncertainties, enhancing temporal fidelity and cross-modal robustness.

By conditioning on the same physical attributes and acquisition settings for both modalities, RISA encodes each scene using the same parameter vector and stochastically transforms the image and point cloud using the same augmentation parameters. Thus, the two sensor streams experience the same physical perturbations while maintaining their sensor-specific degradation signatures. Mapping the shared augmentation parameters to physically-motivated degradation processes specific to images and LiDAR then produces multimodal augmented pairs that are semantically aligned. In this manner, RISA captures how precipitation, illumination, and visibility jointly affect multimodal sensor measurements while maintaining cross-modal consistency. More complex sensor-dependent effects, which are not explicitly modeled by the current implementation, can occur due to the same underlying environmental causes. Some examples include image blur / occlusion caused by raindrops on the lens or LiDAR multipath artifacts induced by high levels of humidity. Modeling these effects would necessitate richer, sensor-level physical models of data acquisition and will be explored as future work. For this reason, RISA should be viewed as a first-order, physics-inspired augmentation technique targeted toward establishing cross-modal consistency, rather than a complete simulator of real-world roadside sensor degradation.

1) Image Branch Augmentation

The image branch adopts an atmospheric scattering-based imaging model as its core mechanism, extended to account for a variety of photo metric degradation factors commonly observed under adverse weather conditions. These include low illumination and overexposure effects during nighttime, as well as visual occlusion caused by rain and snow particles. By integrating these phenomena, the model constructs controlled, composition-preserving degradations that simulate rain, fog, snow, and nighttime environments in a physically plausible manner. The enhancement process for RGB images is formulated as follows:

$$I'(x) = J(x)t(x) + A[1 - t(x)], \quad t(x) = \exp(-\beta d(x)) \quad (2)$$

where $I'(x)$ represents the observed image irradiance at pixel x , $J(x)$ is the scene radiance (clean image), A denotes the global atmospheric light, and $t(x)$ is the transmission map, with β being the atmospheric extinction coefficient derived from the unified scene parameter vector and $d(x)$ the estimated scene depth. In the released implementation, an externally generated dense depth map can be supplied when available, for example from synchronized LiDAR-based preprocessing. When such a depth map is unavailable, RISA uses a deterministic roadside depth prior as a fallback, assigning relatively larger depths to upper image regions and smaller depths to lower image regions. This design preserves the monotonic distance dependence required by the atmospheric-scattering model without relying on a dedicated depth-

estimation network within the RISA module.

To simulate precipitation effects such as rain and snow, RISA applies a combination of forward-view occlusion and scattering-based degradation models. For rain, RISA introduces directional streak-like foreground particles together with atmospheric-scattering degradation. The density and length of the rain streaks are controlled by precipitation-related scene parameters and augmentation severity. The generated streak layer is mildly smoothed and then composited with the original image through alpha blending, thereby modifying local appearance while largely preserving the underlying geometric structure. These image-side rain parameters are coupled with the corresponding LiDAR attenuation and pseudo-echo settings through the shared scene parameter vector, allowing the two modalities to reflect a consistent adverse-weather condition. The streak density and length are jointly modulated with the simulated exposure time, providing consistent control signals that also inform pseudo-echo generation in the LiDAR simulation pipeline (Halder et al., 2019). In contrast, snow is modeled as slow-falling, large-particle occlusions with semi-transparency, causing more localized, discrete blocking effects rather than continuous streaks (Bernuth et al., 2019). Snowflakes are simulated using randomly distributed semi-transparent disks, with particle size and coverage area controlled by a monotonic function of snowfall intensity. Unlike rain, which introduces both visual and LiDAR-level scattering, snow primarily causes occlusion in the image domain and only minor effects on the LiDAR signal—manifesting as slight point dropout and low-intensity noise. Thus, snow does not necessitate heavy pseudo echo modeling on the LiDAR side, maintaining the realism of mild degradation while preserving physical plausibility across modalities.

$$I_{light}(x) = \gamma J(x)^\alpha + \sqrt{J(x)}\varepsilon_{shot}(e) + \varepsilon_{read}(e) \quad \alpha \in (0,1] \quad (3)$$

To simulate low-light conditions such as fog and nighttime, RISA employs an approximate imaging model that incorporates camera response characteristics and noise modeling. Specifically, the model includes parameters for exposure gain γ , sensor electronic noise level e , and gamma compression α . Nighttime degradation is primarily driven by low illumination and sensor noise (Wei et al., 2020). Baseline brightness is adjusted through exposure gain and gamma correction, while Poisson–Gaussian noise is injected to emulate camera-specific signal distortions (Guo et al., 2023). Given that nighttime has minimal impact on LiDAR performance, RISA does not enforce artificial degradation on the point cloud modality in these scenarios. Instead, it maintains strict geometric and temporal alignment to preserve cross-modal consistency without compromising LiDAR fidelity. Fog is characterized by a monotonic reduction in visibility and contrast with increasing distance, along with an upward shift in atmospheric light (Wang et al., 2024b). To capture this effect, RISA increases the extinction coefficient β , and computes the depth-aware transmission map $t(x)$ to simulate the “brighter and grayer in the distance” phenomenon. In parallel, the atmospheric light parameter A is slightly elevated to match the background illumination filled in by scattering. For night-fog scenarios, the fog model is jointly modulated with low-light compensation settings ($\alpha \downarrow, \gamma \uparrow$) to simulate both reduced visibility and exposure adjustments. The resulting image degradation is aligned with LiDAR-based range attenuation using the same underlying scene parameter vector, ensuring physical consistency and synchronized degradation across modalities.

In RISA, the atmospheric-scattering component is explicitly depth dependent, while precipitation overlays, illumination changes, and sensor-noise terms are controlled by the shared scene parameters. This combination introduces distance-dependent visibility degradation while maintaining consistent environmental controls across the image and LiDAR branches. Modeling corruptions in this manner provides one notable benefit: given an environmental condition, augmentations of images and LiDAR share the same underlying parameters. Intuitively, when visibility decreases, atmospheric transmission is reduced and

precipitation particles increasingly occlude image regions, while attenuated returns and sparse pseudo-echoes are introduced on the LiDAR side. Augmentations created in this fashion maintain pixel-semantic consistency with corresponding points in 3D space, allowing multi-modal data to stay physically-consistent and co-interpretable even in degraded states.

2) Point Cloud Branch Augmentation

For LiDAR point cloud data, RISA adopts a physically grounded model based on energy propagation and echo return statistics. Two core variables are used to characterize the degradation process: the point survival probability pulse and the intensity attenuation factor P_r , which describe the likelihood of a point being retained during propagation and the degree of return signal weakening, respectively. These variables are formulated as follows:

$$\begin{aligned} p_{keep}(d) &= \exp(-\kappa\beta d) \\ p_i' &= p_i \exp(-\gamma\beta d) + \varepsilon \end{aligned} \quad (4)$$

Both κ and γ are dynamically modulated by the unified scene parameter vector to ensure consistency with environmental and sensor conditions. Additionally, pseudo-echo points N_{spur} —simulating spurious returns caused by precipitation induced scattering—are sampled from a Poisson distribution, reflecting the stochastic nature of weather-induced LiDAR noise. This formulation enables the generation of physically plausible and statistically coherent point cloud degradations that align with the corresponding image perturbations under the same environmental scenario. To reflect the distinct physical characteristics of various weather conditions, RISA defines condition-specific augmentation strategies for point cloud data. In foggy environments, increased atmospheric scattering leads to distance-dependent signal attenuation and reduced signal-to-noise ratio. As a result, both the point survival probability pulse and the return intensity factor P_r progressively decrease with range. Additionally, slight range jitter is injected to mimic reduced measurement stability at longer distances. This results in sparser, weaker returns for distant objects—consistent with the reduced visibility observed in fog (Teufel et al., 2022). Under rainy conditions, water droplets induce laser scattering and partial reflections, leading to the occurrence of pseudo-echoes (Haider et al., 2023). To simulate this, the pseudo-echo generation process $N_{spur}(r, \lambda)$ is activated, inserting low-intensity false points sparsely along the beam path. When coupled with image-domain rain streaks and atmospheric-scattering effects, this multi-modal augmentation reproduces characteristic appearance degradation together with sparse LiDAR disturbances in rainy scenes. In snowy conditions, snowflakes—being larger, slower, and less reflective—primarily act as occluders rather than scatterers (Hahner et al., 2022). Accordingly, pulse is slightly reduced and mild random noise is applied to the intensity values, without enforcing pseudo-echo generation. While occasional close-range dropout may occur under high snow coverage, the simulation avoids generating dense false points along the beam path, as would be expected in heavy rain. This preserves physical plausibility while minimizing structural distortion in the geometry. Notably, as LiDAR is an active sensing modality, nighttime conditions have minimal direct impact on its measurements (Heinzler et al., 2020). Therefore, in night scenarios, RISA maintains the integrity of the point cloud data and focuses augmentation efforts on the image modality, ensuring modality-specific realism without compromising cross-modal alignment.

For point cloud branch, the physical transmission process from “weather condition $\rightarrow$ atmospheric medium $\rightarrow$ echo response” is modeled into three separately controlled factors - intensity attenuation, point survival rate and pseudo-generated echo respectively. Each degradation term also holds quantitative consistent correlation with counterparts in image branch like transmission, visibility and photometric distortion. This way, we establish a one-to-one relation between two modalities in terms of representing different weather

conditions: foggy translates to range-based sparsification, rainy to motion-blur like jittering with sparse pseudo-generated echo, snowy to occlusion dominant slight point dropout, night to no obvious effect on LiDAR modality. As a result, the sensor response between modalities become cross-consistent and easily interpretable when facing complex weather disturbances.

3) Augmentation Strategy Archiving

Five discrete levels of augmentation severity, denoted as C1 through C5, are defined based on the multi-modal data augmentation applied to both image and point cloud branches, corresponding to an increasingly difficult curriculum of environmental conditions. C1 represents nominal conditions with minimal degradation (e.g., light noise), while C5 simulates extremely adverse scenarios such as dense nighttime fog and heavy rain. Intermediate levels (e.g., C3) involve moderate disturbances, such as mild fog or light rain, allowing for a progressive increase in task difficulty.

To quantitatively describe the consistency of data augmentation across the image and point cloud modalities, this paper introduces two metrics: re-projection error and intersection over union (IoU). The re-projection error is calculated by projecting LiDAR points onto the camera image plane and measuring the alignment between the projected points and the corresponding image features. The re-projection error is computed using the following formula:

$$E_{reproj} = \frac{1}{N} \sum_{i=1}^N \|P_i - I_i\| \quad (6)$$

where N is the number of LiDAR points, P_i is the projection of the i -th LiDAR point onto the image plane, and I_i is the position of the corresponding feature in the image.

Additionally, this paper evaluates the intersection over union (IoU) between LiDAR and image features in the bird's-eye view (BEV) space to further quantify the degree of spatial alignment. The IoU is calculated by measuring the ratio of the overlapping area between the LiDAR point cloud and the image features to the total area in BEV space, as follows:

$$IoU = \frac{\text{Area of intersection}}{\text{Area of Union}} \quad (7)$$

To enable precise control and reproducibility of the augmentation process, RISA logs augmentation metadata in a structured format. This metadata includes not only the applied parameter values for each augmented sample pair, but also the associated difficulty level of the scene. By archiving this information, a traceable record of all multi-modal augmentation strategies is established, supporting curriculum-aware training workflows and enabling historical introspection for future curriculum updates.

3.2. SSL Coupling Network

In each training iteration, the system takes as input a paired set of synchronized samples: the original data (I, P) and its corresponding RISA-generated augmented version (I', P'). These pairs are processed through shared-weight image and point cloud encoders, which extract modality-specific features for both detection and representation learning. The resulting features are then aligned into a unified representation space via a lightweight fusion neck.

A teacher-student dual-network architecture is adopted to enhance robustness without requiring additional annotations. The teacher network receives the original (clean) samples and maintains stable predictions via momentum-based updates using exponential moving average (EMA). The student network ingests the augmented samples and learns through gradient based back propagation. This design encourages the model to produce consistent predictions across different views of the same scene—original versus perturbed—thereby improving cross-modal feature alignment and robustness against perturbations.

The overall training objective consists of three components: a supervised detection loss L_{det} , a cross-modal contrastive loss L_{cm} , and a teacher-student consistency loss L_{cons} . The combined objective function is defined by

$$L = L_{det} + \lambda_{cm} L_{cm} + \lambda_{cons} L_{cons} \quad (8)$$

where λ_{cm} and λ_{cons} are weighting factors dynamically adjusted by the curriculum strategy, which gradually ramps up the influence of the self-supervised losses as training progresses. This staged introduction prevents early overfitting to noisy augmentations and guides the model toward learning robust, semantically aligned representations under increasingly challenging conditions. As illustrated in Figure 2, the architecture of the SSL Coupling Network is composed of modality-specific encoders, a geometric-semantic fusion neck, and a 3D detection head. On the image side, the encoder adopts a convolutional backbone with a feature pyramid network (FPN) to extract multi-scale semantic features. On the LiDAR side, the released configuration employs dynamic voxel feature encoding followed by sparse 3D convolution and a SECOND-style BEV backbone to extract geometric representations. The resulting image and LiDAR features are subsequently mapped to the common fusion space for geometric and semantic alignment. The fusion neck comprises two key stages.

- 1) Geometric projection alignment: Using the Camera-LiDAR calibration parameters, BEV grid centers are projected onto the image plane. Corresponding image features are then sampled at the projected locations, ensuring that the two modalities are aligned within a shared spatial unit.
- 2) Semantic pooling and lightweight projection: Within candidate regions or semantic masks, cross-modal pooling is performed, followed by a lightweight projection head that maps features into a unified embedding space. This aligned space facilitates cross-modal contrastive learning and consistency supervision. The final 3D detection head outputs bounding boxes and category distributions.

Unlike typical pixel-level or BEV-level loose alignments, our method leverages explicit geometric projection combined with semantic-region aggregation. The former ensures accurate spatial correspondence, while the latter preserves semantic object integrity. Together, they establish a stable basis for both cross-modal and cross-augmentation matching (Li et al., 2026).

Effective contrastive learning requires high-quality positive samples (Choo et al., 2024). A set of reliable pixel-BEV correspondences is constructed, constrained by calibration accuracy and visibility. Only correspondences with consistent depth, no occlusion, and minimal projection error are retained as equivalent expressions of the same object across modalities. Hard negatives are sampled from spatially adjacent but semantically dissimilar units, as well as from a momentum-based memory queue across batches. This sample construction strategy mitigates noise from occlusions, misprojections, and pseudo-echoes, ensuring that contrastive objectives promote truly geometry-semantic aligned representations (Cai et al., 2024).

1) Cross-modal constraint

multi-modal sensory data inherently exhibit different physical properties at the time of acquisition. To minimize feature misalignment caused by modality-specific discrepancies during training, contrastive constraints are imposed, encouraging representations of the same real-world object across modalities to be similar, while separating representations of different objects or instances from different time steps. These constraints are applied at both global and local levels of granularity.

At the global level, global average pooling is performed over the entire image and LiDAR frames to obtain modality specific global semantic vectors. F_i^{img} and F_i^{lid} denote the image and point cloud

features extracted at the i -th scale, and η_i represent the full multi-scale feature sequences from the respective feature pyramid networks. The global semantic embeddings are then computed as:

$$g_{img} = \sum \eta_i Pool(F_i^{img}) \quad (9)$$

$$g_{lid} = \sum \eta_i Pool(F_i^{lid}) \quad (10)$$

A global cross-modal contrastive loss is introduced to enforce global-level semantic alignment, defined as:

$$L_{global} = -\log \frac{\exp(\text{sim}(g_{img}, g_{lid})/\mu)}{\sum_{q \in Q} \exp(\text{sim}(g_{img}, q)/\mu)} \quad (11)$$

where sim denotes the cosine similarity (normalized dot product), μ is the temperature coefficient, and Q is a momentum queue containing negative samples from previous batches. This loss encourages alignment of image and point cloud global features at the scene level, thereby mitigating modality gaps in the shared semantic space.

At the local level, a set of reliable cross-modal correspondences is constructed during feature extraction by leveraging sensor calibration to project LiDAR pillar units onto the image plane. Let P denote the set of 3D pillars and u the corresponding set of projected image pixels. For each projected pair, local feature vectors f_p^{lid} and f_u^{img} are extracted, enabling the establishment of object-level and geometric-level consistency between modalities. To enhance discriminability, hard negative samples are preferentially selected from spatially adjacent but semantically distinct pillar-pixel pairs, in addition to random negatives drawn from the momentum queue. The local cross-modal contrastive loss is defined as:

$$L_{local} = -\log \frac{\exp(\text{sim}(f_p^{lid}, f_u^{img})/\mu)}{\sum_{q \in N} \exp(\text{sim}(f_p^{lid}, q)/\mu)} \quad (12)$$

where N denotes the set of negative samples, including inframe semantically mismatched pixel-pillar pairs and historical negatives from previous batches. This loss constrains the local semantic alignment between modalities and complements the global alignment objective. Finally, the total cross-modal contrastive loss combines both levels of supervision:

$$L_{cm} = \alpha_{global} \cdot L_{global} + \alpha_{local} \cdot L_{local} \quad (13)$$

The weights α_{global} and α_{local} are dynamically updated based on the curriculum strategy throughout training. This global + local design effectively avoids semantic drift that may arise from purely local alignment, while compensating for the lack of spatial precision in global-only constraints—thus achieving robust and fine-grained cross-modal representation learning.

2) Teacher-Student Consistency

The core assumption of consistency learning is that data augmentation should not alter semantic meaning — that is, the augmented sample should preserve both the semantic and geometric content of the original. Based on this assumption, the student network is constrained to replicate the teacher network's behavior at two levels: feature representation and classification output. F_T and F_S denote the intermediate features extracted from the teacher and student networks, respectively.

Feature Consistency: Given that RISA explicitly records the geometric and photometric transformations applied during augmentation, the teacher's intermediate feature maps are reprojected into the coordinate system of the student's view. The difference between aligned features is then measured using the mean squared error (MSE), encouraging invariance to degradation effects such as rain, fog, low light, and random noise. The feature-level consistency loss is computed as:

$$L_{feat} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\phi F_T(x) - \phi F_S(x)\|^2 \quad (14)$$

where Ω denotes the set of valid positions after spatial alignment, and ϕ represents the

geometric-photometric warping function that reprojects teacher features into the student's spatial frame using known augmentation parameters. This constraint ensures that feature representations remain

consistent across perturbations, thereby improving the model's stability and generalization under challenging environmental conditions.

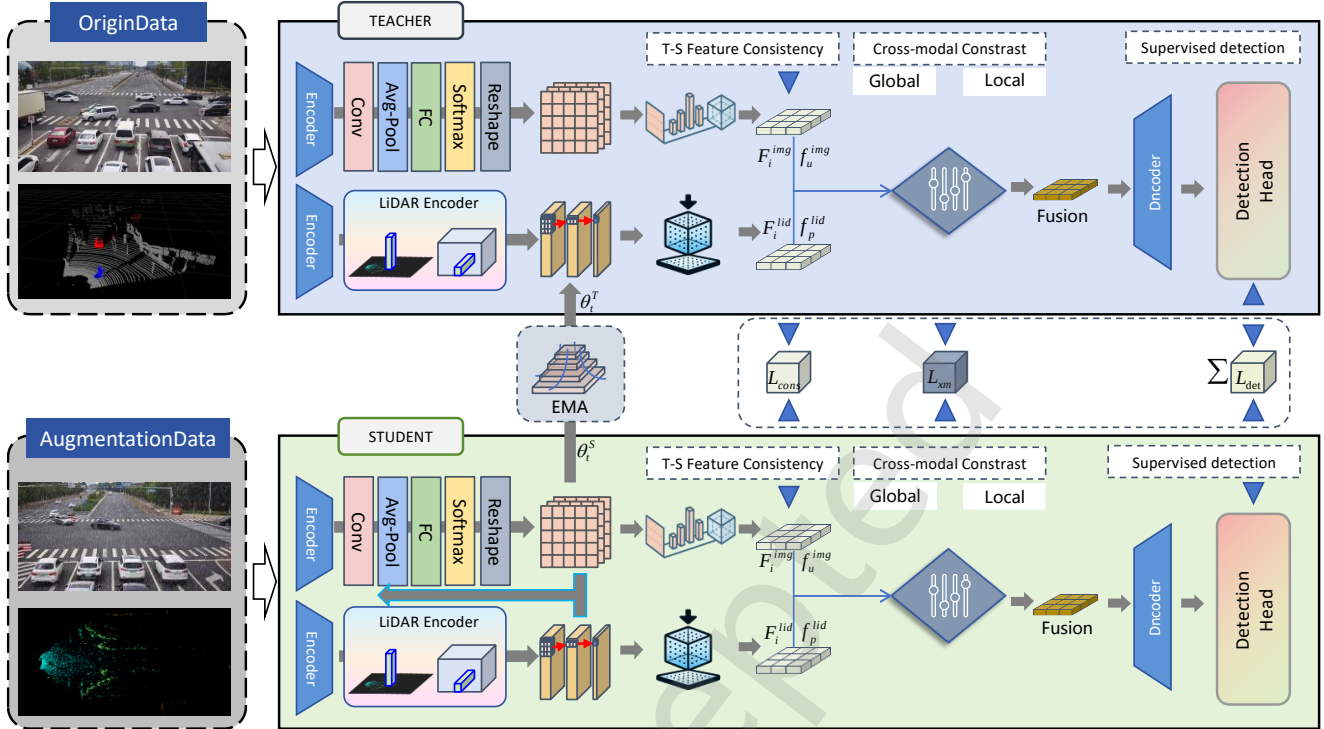


Fig. 2. Coupled network architecture based on self-supervised learning. This diagram illustrates the coupled network architecture, consisting of a teacher-student framework, where the teacher network processes clean data, and the student network learns from augmented data. The teacher-student consistency and cross-modal contrastive losses are combined to enhance model robustness across modalities.

Classification Consistency: For the aligned valid regions between teacher and student views, the Kullback-Leibler (KL) divergence between their predicted class distributions with temperature scaling is computed. To prevent the student from learning from noisy or uncertain signals, only samples where the teacher's confidence exceeds a predefined threshold are included. Additionally, a soft weighting mechanism based on the teacher's confidence is applied to further suppress the influence of potentially unreliable pseudo-labels. This approach leverages the stability of the teacher while avoiding the propagation of incorrect supervision to the student. The classification-level consistency loss is defined as:

$$L_{cls} = \sum_{x \in \Omega} \omega_x \text{KL}(\sigma(s_x^T) \parallel \sigma(s_x^S)) \quad (15)$$

where s^T and s^S denote the class logits predicted by the teacher and student at position x , σ is the softmax function, and $\text{KL}(\cdot \parallel \cdot)$ represents the Kullback-Leibler divergence. The confidence weight ω is computed based on the teacher's softmax confidence. The overall teacher-student consistency loss combines both feature-level and classification-level components:

$$L_{cons} = \sum (L_{feat}, L_{cls}) \quad (16)$$

While feature consistency promotes stability in intermediate representations under augmentation, classification consistency reinforces semantic label in variance. Together, these constraints guide the student network to learn robust and consistent representations across perturbations, significantly improving model generalization in complex environments.

The teacher network serves as a stable reference for the student, providing consistent feature representations and classification predictions. Unlike the student network, the teacher is not updated through back propagation. Instead, its parameters are inherited via Exponential Moving Average (EMA) updates from the student

network. The update rule is defined by

$$\theta_t^T = m \cdot \theta_{t-1}^T + (1 - m) \cdot \theta_t^S \quad (17)$$

where θ_t^T and θ_t^S denote the teacher and student network parameters at iteration t , respectively, and $m \in [0, 1]$ is the momentum coefficient. This EMA-based update mechanism allows the teacher to smoothly accumulate the student's parameters over a longer time horizon, effectively filtering out stochastic noise and unstable gradients from individual iterations. As a result, the teacher network produces more robust and temporally stable outputs, which serve as reliable supervision signals for enforcing consistency on augmented data.

Conceptually, the teacher can be viewed as a momentum-averaged version of the student throughout training, providing a denoised and stabilized prediction trajectory. This design enhances the quality of consistency regularization and plays a critical role in ensuring that the student learns invariant representations aligned with unperturbed semantic content.

3.3. VLM Curriculum Scheduler

The VLM scheduler is an outer-loop module that takes structured feedback from training processes (e.g. loss statistics, alignment metrics, representative failure cases), and produces a runnable curriculum policy (written in YAML) that prescribes how sampling/augmentation should be changed for next round of training. We treat this module as the outer-loop controller of cross-iteration training that translates structured training feedback to runnable curriculum policies. Under joint physical and semantic constraints, it continuously expands the model's effective training domain. At the end of each training round, the system aggregates a structured data package that includes: detection loss and recall metrics segmented by scenario and difficulty level; cross-modal alignment indicators such as contrastive agreement rate and BEV

projection deviation; teacher-student divergence and teacher confidence histograms; statistics on augmentation strength and sampling frequency; and a small set of representative failure cases with paired image-point cloud snapshots and textual summaries. These numerical, visual, and textual signals are assembled into a template-based prompt, which is fed into the VLM scheduler. Guided by embedded physical priors and task-oriented instructions, the VLM scheduler generates interpretable natural language decisions on: which scenarios remain underrepresented, which augmentation combinations are physically invalid, and which threshold settings are too lenient or too strict. These decisions are deterministically parsed into structured curriculum strategies π_{t+1} in YAML format, ensuring reproducibility and traceability of all scheduling decisions.

The VLM scheduler adheres to three core principles: physical consistency, semantic robustness, and computational feasibility. The strategy generation process of the VLM scheduler is governed by three guiding principles: physical consistency, semantic robustness, and computational feasibility.

At the physical level, hard constraints are applied to each candidate augmentation parameter vector s to enforce cross-modal compatibility $h_{phys}(s) \leq 0$. For example, an increase in rain intensity on the image side must be accompanied by consistent adjustments to LiDAR point survival rates and pseudo-echo probability as a function of range; in pure nighttime scenes, LiDAR retention should remain unaffected.

At the semantic level, the supervisor evaluates sample quality through a gating function $q(x) = \rho_{cm}(x) \geq \tau_{cm} \wedge \text{conf}_T(x) \geq \tau_{conf} \wedge \kappa(s) \leq \tau_{phys}$ based on teacher confidence and cross-modal consistency. Only samples meeting a quality threshold are allowed to participate in self-supervised loss computation, thus preventing unreliable pseudo-labels and semantically incoherent perturbations from degrading model performance.

At the computational level, the scheduler dynamically balances the benefits of difficulty escalation against the associated computational cost and convergence risks, yielding an adaptive and efficient curriculum for the next training round.

$$\pi_{t+1} = \arg \min_{\pi \in \Pi} E_{s \sim \pi} [R_t(s)] + \mu C(\pi) \quad s.t. \quad h_{phys}(s) \leq 0 \quad (18)$$

The sample quality function $R_t(s) = \alpha L_{det}^{val_T-s_{cm}}$ aggregates multiple feedback signals—validation-set detection risk, teacher-student divergence, and cross-modal alignment metrics—to estimate sample hardness and semantic stability for different scene types. $C(\pi)$ is the estimation of statistical items based on scene difficulty and sample ratio. Rather than being optimized through back propagation, $R_t(s)$ is produced by the VLM scheduler via a combination of heuristic rules and model-based reasoning, resulting in a feasible and feedback-responsive discrete sampling distribution and augmentation parameter ranges that reflect the model's evolving needs. In implementation, the VLM combines structured feedback with predefined heuristic and semantic reasoning rules to produce a candidate preference over discrete scene - difficulty buckets. This candidate is subsequently converted into an executable sampling policy through the deterministic procedure described in Eqs. (21) - (23). Therefore, Eq. (18) specifies the desired decision principle of the outer loop, whereas the actual detector parameters remain optimized only by the supervised and self-supervised objectives defined in the inner loop.

To ensure that curriculum evolution remains stable and controllable, the VLM scheduler employs an update policy based on difficulty ramp-up with weak-scene reweighting. The augmented domain is discretized into K scene-difficulty buckets (e.g., rain/snow/fog/night $\times$ C1-C5), each associated with a sampling probability $p^t \in \Delta^{K-1}$. These probabilities are multiplicatively updated according to the relative improvement-to-cost ratio Δu_k^t observed in the most recent training rounds:

$$p_k^{t+1} \propto p_k^t \exp(\eta \Delta u_k^t) \quad (19)$$

$$\Delta u_k^t = \text{norm}[mAP_k^t - mAP_k^{t-1}] - \lambda_{stab} \text{norm}[\delta_{T-S}, k] \quad (20)$$

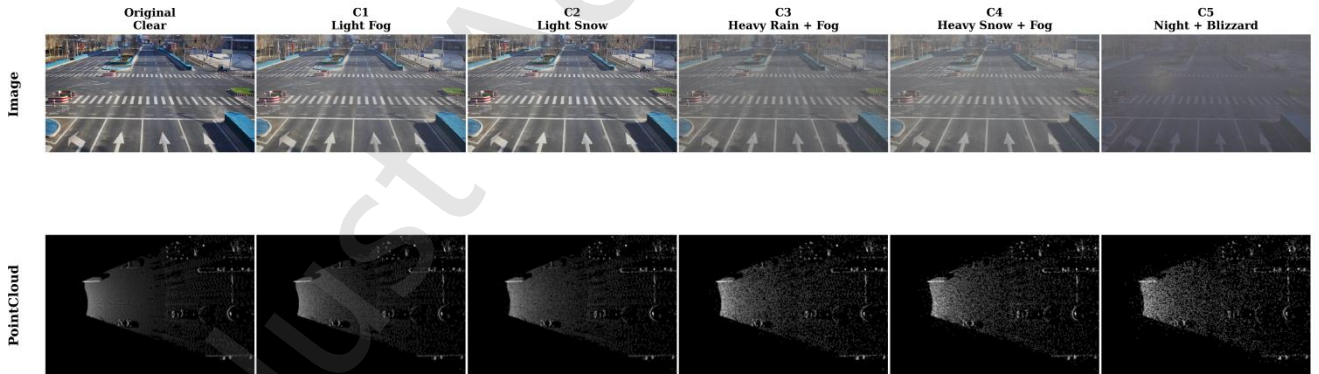


Fig. 3. Data augmentation data enhanced by RISA. Examples of data augmentation using the RISA module, showing how different levels of environmental perturbations (e.g., rain, fog, snow, and nighttime) are applied to both image and LiDAR data. This figure illustrates the effects of augmentation on both data modalities at varying levels of intensity (C1 to C5).

We keep sampling from a small subset of low-difficulty samples during training to achieve algorithmic stability and avoid catastrophic forgetting, while the rest of the sampling distribution are adjusted by bias and scaling factor continuously learned from recent failure cases. For example, if failure cases are biased towards long-range detection under fog condition, we scale up the sampling probabilities from buckets corresponding to fog condition in VLM scheduler, as well as tighten up the high end threshold of transmission rate in modality imbalance evaluator on image side for stronger depth-aware contrast degradation recovery. Similarly, if failures cluster in high-severity rain while teacher-student disagreement indicates excessive LiDAR corruption, the scheduler may increase the image-side precipitation strength while constraining the pseudo-echo rate and raising the teacher-confidence threshold. This avoids unnecessarily strong LiDAR degradation while retaining sufficiently challenging image-side

perturbations. With the output from VLM scheduler taking charge of the RISA sampling as well as SSL coupling in the next iteration, our training loops become fully reproducible, traceable, and adaptively optimized. VLM scheduler is trained to be a physically bounded policy adjudicator instead of a black-box auto-tuner. On one hand, we break down its decision-making process with multi-modal evidence and post-hoc explainable, natural language based rationales on why the model failed, where it failed and how the augmentation strategy should be further adapted. On the other hand, its language output is parsed into structured strategy templates that are subjected to strict physical constraint validation and offline replay verification to prevent undesirable strategy drift caused by the stochastic nature of language models.

To avoid treating the VLM scheduler as a direct optimizer of detector

parameters, ROSE models the VLM scheduler update as a constrained curriculum-policy update over predefined scenario buckets. Let the training set be partitioned into K curriculum buckets according to scenario type, weather condition, and augmentation difficulty. The sampling policy at round t is denoted by $\pi_t \in \Delta^K$, where Δ^K is the K -dimensional probability simplex.

Given aggregated training feedback, the VLM scheduler output is first parsed into a structured score vector $z_t \in \mathbb{R}^K$ after prompt interpretation, physical-validity checking, and semantic-consistency checking. The candidate policy induced by the VLM scheduler is then defined as

$$\tilde{\pi}_{t+1} = \text{Softmax}(z_t/\delta) \quad (21)$$

where $\delta > 0$ is a temperature parameter controlling the sharpness of the bucket preference distribution.

To suppress abrupt policy oscillation across consecutive rounds, the final curriculum update is implemented using a damped projection:

$$\pi_{t+1} = \Pi_{\Delta}(1 - \xi)\pi_t + \xi\tilde{\pi}_{t+1} \quad (22)$$

where $\xi \in (0, 1]$ is the update damping factor and Π_{Δ} denotes projection onto the probability simplex. This formulation guarantees that π_{t+1} remains a valid probability distribution. Moreover, since both π_t and $\tilde{\pi}_{t+1}$ lie in Δ^K , the one-step policy variation is bounded by

$$\|\pi_{t+1} - \pi_t\| < 2\xi \quad (23)$$

which limits abrupt curriculum fluctuation and provides a bounded-update constraint for outer-loop scheduling. Therefore, the theoretical support provided here concerns the validity and boundedness of curriculum evolution, rather than a strict global convergence proof for the entire non-convex detector training process. It should also be noted that this bounded update formulation limits the one-step variation after a candidate policy has been generated, but it does not completely eliminate the stochasticity of the VLM-derived score vector z_t . Since z_t may still be affected by prompt design and model decoding randomness, large fluctuations in the parsed score vector could potentially lead to policy oscillation in long-term deployment. To mitigate this risk, ROSE adopts deterministic prompt templates, structured YAML parsing, physical-validity checking, semantic-consistency checking, and damped projection. In the present experiments, the random-seed settings for Random Curriculum, Offline-VLM, and Online-VLM are fixed and explicitly recorded. For each strategy, performance is accumulated and averaged over multiple scheduling rounds under the corresponding random setting. This round-level aggregation reduces the influence of any single stochastic scheduling event and supports the stability of the observed performance differences. A broader independent-retraining study across additional seeds and prompt variants will be considered in future work.

Note that under this recipe, the VLM scheduler only participates in data selection and difficulty-paced sampling for next-round training, while detector parameters are exclusively updated by supervised-loss & self-supervised consistency objectives as before. The scheduler also blocks updates that violate physics rules or fail semantic-consistency checks, falling back to the last accepted policy or a conservative policy update if such issues are encountered.

End-to-end, curriculum scheduling is no longer static priors but adaptive heuristics contingent on the learning-state of the model. For instance, once the detector stabilizes on a certain class of difficult samples, the scheduler can ease the curriculum by sampling from higher difficulty intervals or assigning higher probability to rare-instance scenarios. On the other hand, if training enters an unstable regime with signs of error amplification such as growing self-supervised losses, worse cross-modal consistency, or uneven validation accuracy, the scheduler can react by freezing the curriculum difficulty, tightening sample-quality filters, and scaling down teacher-student consistency supervision to prevent contamination from noisy pseudo-

labels. Empirically we found this mechanism led to smooth gains in robustness over training rounds with less noisy loss fluctuations.

Since the VLM scheduler in ROSE only processes aggregated metadata from training batches and is kept separate from gradient computations, the cost of VLM inference is decoupled from the inner-training loop that updates detector weights. Thus additional VLM-related latency does not bottleneck the critical back-prop path in every training iteration and only influences training throughput. In our current implementation, the cloud-based VLM API is used as a practical realization to validate the proposed feedback-driven curriculum scheduling mechanism. Nevertheless, when ROSE is extended to edge-side continual learning or long-term adaptive training, repeated access to a cloud-based VLM may introduce additional communication latency, service dependency, and monetary cost. To improve practical deployability, several engineering-oriented variants can be considered. First, historical training-feedback and curriculum-policy pairs can be cached, allowing previously validated scheduling strategies to be reused through log-replay-assisted matching when similar training states are encountered. Second, the VLM scheduler can be invoked selectively rather than periodically, for example only when instability indicators such as increased teacher-student divergence, degraded cross-modal consistency, or concentrated failures in weak scenarios exceed predefined thresholds. Third, asynchronous policy generation can be adopted to prevent VLM calls from blocking the main training process. In addition, accumulated VLM-generated policies may be distilled into a lightweight local scheduler or converted into rule-based fallback strategies for edge-side execution. These deployment variants indicate that the online cloud-based VLM scheduler is not the only feasible deployment form. When network latency, API cost, or service stability becomes a constraint, an offline/cached scheduler or a lightweight local scheduler can serve as a more practical trade-off for edge-side continual learning. Since these variants mainly concern engineering deployment rather than the core algorithmic design of ROSE, they are not systematically evaluated in the present study. A more comprehensive study of offline caching, selective invocation, asynchronous policy generation, and lightweight local scheduling will be explored in future work.

4 Experiments

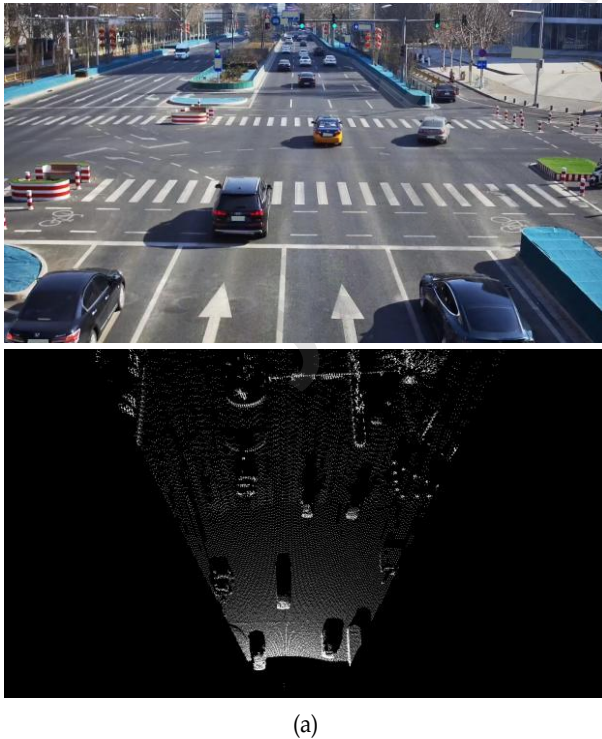
4.1. Experimental Setup

To comprehensively validate the effectiveness of the proposed ROSE framework for roadside multi-modal perception enhancement and self-supervised coupling, we design and conduct systematic experiments. The DAIR-V2X-I roadside multi-modal dataset is adopted as the foundational benchmark (Yu et al., 2022). To verify the cross-data generalization ability of the ROSE framework, the performance on the RCooper roadside dataset was compared (Hao et al., 2024). It should be noted that the use of DAIR-V2X-I and RCooper is not a simple repeated evaluation on two similar benchmarks, but an initial cross-dataset validation under different roadside deployment settings. DAIR-V2X-I is built on real-world vehicle-infrastructure scenarios collected in the Beijing high-level autonomous driving demonstration area, covering multiple intersections, city roads, and highway segments. Its infrastructure side mainly uses a 300-beam LiDAR and an RGB camera as the roadside sensing pair. In contrast, RCooper focuses on practical roadside cooperative perception and contains corridor- and intersection-oriented deployment schemes with more heterogeneous infrastructure sensor combinations, including 80-beam and 32-beam LiDAR groups, MEMS LiDAR, and RGB cameras. These differences lead to variations in collection region, roadside sensing coverage, installation layout, and hardware configuration. The comparative verification based on two public datasets can provide a representative preliminary illustration of the cross-dataset adaptability of ROSE in different real roadside scenarios.

Figure 4 presents typical road scenes from the DAIR-V2X-I and

RCooper datasets. The upper panel shows RGB images captured by roadside cameras at a frequency of 25 Hz, while the lower panel displays the corresponding laser point cloud images captured simultaneously at a frequency of 10 Hz. These paired data modalities form the basis of our multi-modal perception and enhancement experiments. Since the DAIR-V2X-I and RCooper datasets mainly contain daily urban traffic scenes and are difficult to cover a large number of special scenarios (such as rain, snow, and fog), this paper enhanced the original datasets based on the initial curriculum strategy. Specifically, the RISA enhancement module is used to generate cross-modal paired enhanced samples that conform to a unified scene parameter vector control and have realistic physical effects. To deeply analyze the performance of the model under different levels of perception interference, as shown in Figure 3, multiple enhancement intensity levels were set for the experiments to explore the performance change trends and robustness boundaries of the proposed method under controllable enhancement conditions. Figure 5 shows the initial curriculum strategy.

All experiments are performed on a workstation with AMD Ryzen 9 79 45HX w/Radeon Graphics (2.50GHz) and NVIDIA RTX 40 80S GPU. For VLM-assisted curriculum scheduling, we deploy Qwen as the external supervisory model through the Alibaba Cloud platform. The scheduler is invoked at predefined curriculum-update intervals, with approximately six invocations during each complete training cycle. It receives summarized training dynamics rather than raw sensor data. Inputs provided to the scheduler consist of per bucket validation performance, teacher-student disagreement and cross-modal consistency, as well as summaries of recent failures. The outputs of the scheduler include bucket-level curriculum updates: sampling-weight updates, difficulty pacing adjustments, and anomaly threshold adaption. Detector parameters continue to be trained with their original supervised and self-supervised objectives. Detection performance is reported following the 3D object detection protocol of the KITTI benchmark.



(a)



(b)

Fig. 4. Roadside multi-modal scenario data. (a) DAIR-V2X-I dataset (b) RCooper dataset (Up: Camera image; Down: Synchronized LiDAR point cloud visualization)

4.2. Performance Comparison and Analysis

A series of comparative experiments are conducted on the DAIR-V2X-I and RCooper datasets to thoroughly validate the perception capabilities of the proposed ROSE framework for roadside multi-modal enhancement and self-supervised learning. Table 1 summarizes the performance of our approach versus several baseline models in terms of detection accuracy and inference performance.

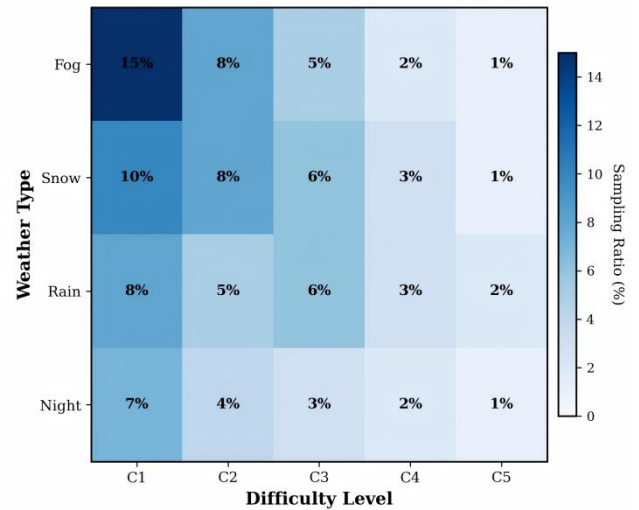


Fig. 5. Initial curriculum strategy. The figure illustrates the distribution of difficulty levels and the proportion of weather types in the data augmentation strategy used during the first epoch of training.

As shown in Table 1, the proposed ROSE framework delivers competitive overall detection accuracy on both the DAIR-V2X-I and RCooper datasets, while maintaining real-time inference speed and balanced performance across multiple object categories. Although

ROSE is not the top-performing method on every individual metric, it exhibits a favorable accuracy-efficiency trade-off and stable generalization behavior under multi-modal adverse-weather settings. These results indicate that the proposed physically plausible augmentation and self-supervised coupling strategy can improve robustness without introducing prohibitive deployment cost. To provide a clearer view of computational cost, we report the standalone computational profile of ROSE. Under the current input setting, ROSE contains 58.3M trainable parameters and requires 236 GFLOPs per frame. On an NVIDIA RTX 4080S GPU, it achieves an end-to-end inference latency of 72-75 ms, with an average GPU utilization of approximately 64% and a peak memory footprint of about 8.4 GB. On an AMD Ryzen 9 7945HX CPU, preprocessing and data packaging require about 5.6ms per frame, with an average CPU utilization of approximately 24% during online inference. We further report the training-time overhead of the VLM scheduler separately from inference latency. The scheduler is invoked approximately six times in each training cycle, with an average API response time of about 1.2 s per call. Considering the accumulated calls over multiple training cycles, together with response parsing, policy validation, and synchronization with the training pipeline, the total wall-clock training time increases from 20.0 h to 21.2 h. This additional cost is measurable but remains controllable. More importantly, the VLM scheduler is used only during the training or curriculum-updating stage and is not involved in inference-time computation. Therefore, its overhead mainly affects training efficiency rather than online deployment latency.

In addition to the metrics above, we evaluate cross-modal fusion consistency on the RISA-augmented test set using Fusion Consistency Rate (FCR). Concretely, FCR is calculated by measuring the cosine similarity between the semantic feature vectors of matched pillar voxels in bird's-eye-view (BEV). Then all matched similarities are averaged to produce a final score that quantifies alignment of spatial semantics between image and LiDAR branches. Retention Rate is used to describe the robustness as perception quality drops. It calculates how much detection accuracy is retained when facing bad conditions including rain, snow and fog. Therefore, Retention Rate provides an explicit measure of the model's capability to retain its performance when confronted with harsh sensing conditions.

To evaluate the applicability of ROSE in roadside scenarios, we further conduct a fine-grained analysis from two perspectives, namely target distance and occlusion severity, as shown in Table 2. From the distance perspective, ROSE maintains relatively stable detection performance across near-range (0-20m), mid-range (20-30m), and far-range (>30m) targets. Meanwhile, the performance degradation under far-range conditions is controlled within 6.6% and 6.3%, respectively,

which is better than that of most competing methods. In particular, ROSE exhibits a noticeably milder degradation trend under mid and far range conditions, indicating that it can still preserve strong cross-modal representation capability when target scale decreases, point clouds become sparse, and visual details are weakened. This property is especially important for roadside early warning and cooperative perception, since medium and long range targets often determine whether the system can achieve earlier risk recognition and response.

From the occlusion perspective, ROSE also demonstrates stronger robustness. Under light occlusion (occlusion ratio<15%), moderate occlusion (occlusion ratio≤30%), and severe occlusion (occlusion ratio≤50%), ROSE shows a more gradual performance degradation trend than the competing methods. This indicates that the proposed physically plausible augmentation mechanism and cross-modal self-supervised coupling strategy are able to preserve more reliable semantic and geometric cues when targets are partially occluded, visibility is degraded, and local information is missing, thereby improving detection stability in complex roadside scenarios.

These results provide a useful complement to the overall mAP, FCR, and Retention Rate metrics, and further offer a more complete validation of the practical value of the proposed method in real roadside deployment scenarios.

As illustrated in Figure 6, increasing the augmentation severity via RISA leads to varying degrees of degradation in both detection accuracy and Fusion Consistency Rate (FCR) across all evaluated models. Notably, ROSE consistently occupies the upper-right region of the FCR-mAP plots across all augmentation levels, indicating its dual advantage in modality alignment quality and perceptual accuracy. Compared to other methods, ROSE shows a more concentrated and steeper performance curve, suggesting lower sensitivity to fusion degradation and higher consistency under perturbations. This implies that the model maintains robust feature alignment and detection performance even under increasingly challenging conditions. Furthermore, in terms of robustness evaluation, ROSE achieves a Retention Rate of 87.5%, significantly outperforming all competing baselines. This result confirms ROSE's strong environmental adaptability and resilience, as it maintains high perception accuracy under adverse weather conditions such as rain, snow, and fog, with minimal performance drop. It should be noted that Figure 6 is mainly used to evaluate the robustness performance of different methods when they encounter progressively enhanced RISA disturbances during the testing phase. To further verify whether ROSE still maintains its advantage when the baseline method also uses RISA for enhanced training, we have added a RISA-augmented baseline training control experiment in Section 4.3.

Table 1. Overall performance comparison

Method	Detection Accuracy (%)								Inference Performance	
	DAIR-V2X-I				RCooper					
	Overall	Car	Cyclist	Pedestrian	Overall	Car	Cyclist	Pedestrian	FPS	Delay(ms)
TransFusion (Bai et al., 2022)	71.3	84.6	60.2	69.1	69.5	82.5	58.6	67.5	11.1	90
BEVFusion (Liang et al., 2023)	73.5	86.9	62.1	71.5	72.3	86.4	61.5	69.0	15.0	67
AutoAlignV2 (Chen et al., 2022)	75.1	87.3	62.7	75.3	77.1	89.3	64.5	77.5	13.8	72
SupFusion (Qin et al., 2023)	77.0	87.8	64.4	78.8	77.9	88.8	65.2	79.6	12.0	83
ECFusion (Fu et al., 2024)	80.9	90.2	69.7	82.8	79.4	89.8	68.9	79.4	14.1	71
ROSE	78.8	89.9	67.9	78.6	78.1	89.6	67.3	77.5	12.0-15.0	72-75

Table 2. Distance- and occlusion-stratified performance comparison of ROSE and representative baselines

Method	Detection Accuracy (%)											
	DAIR-V2X-I						RCooper					
	Near AP (0-20m)	Mid Drop (20-30m)	Far Drop (>30m)	Easy AP	Moderate Drop	Hard Drop	Near AP (0-20m)	Mid Drop (20-30m)	Far Drop (>30m)	Easy AP	Moderate Drop	Hard Drop
TransFusion	74.2	-6.7	-10.4	73.0	-2.2	-4.1	72.1	-6.0	-9.3	71.7	-3.4	-6.2
BEVFusion	76.6	-7.1	-11.2	75.3	-2.2	-4.6	75.2	-6.7	-9.6	73.7	-2.8	-5.4
AutoAlignV2	77.8	-6.2	-9.7	77.1	-2.7	-4.6	79.7	-5.9	-9.3	77.8	-3.4	-5.7
SupFusion	79.5	-5.7	-9.0	78.6	-2.0	-4.0	80.2	-5.2	-8.6	81.6	-3.8	-6.4
ECFusion	83.6	-6.3	-9.7	82.8	-2.4	-4.7	82.1	-6.5	-8.8	84.9	-2.7	-5.1
ROSE	80.8	-3.4	-6.6	80.1	-1.7	-3.1	79.9	-4.0	-6.3	80.4	-2.3	-4.4

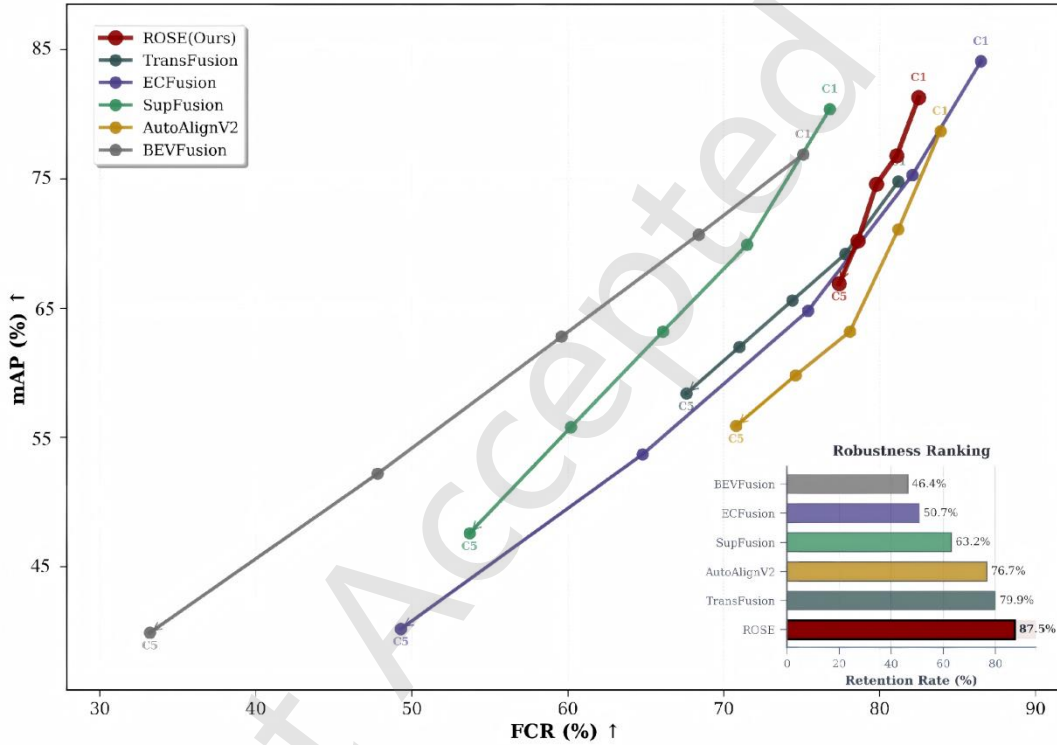


Fig. 6. Robustness comparison across models under progressively stronger RISA augmentations (mAP vs. FCR). Across all augmentation levels, ROSE consistently achieves higher mAP and higher FCR (occupying the upper-right region), and also attains the best performance on the robustness retention-rate metric. These results indicate that ROSE provides stronger robustness under adverse-weather perturbations and more reliable cross-modal semantic alignment.

4.3. Ablation Study

Beyond its engineering reproducibility and inference efficiency, ROSE demonstrates superior robustness and multi-modal alignment quality. These performance gains primarily stem from three core components: (i) a physically plausible cross-modal augmentation strategy (RISA); (ii) the integration of cross-modal self-supervised learning (SSL), and (iii) the VLM scheduler for dynamic curriculum adaptation. To investigate the individual contribution of each module, a series of ablation experiments are conducted on the RISA augmentation module, the SSL coupling network, and the VLM scheduler.

1) Ablation of the RISA Augmentation Module

To evaluate the impact of the RISA cross-modal augmentation module, controlled experiments are conducted to analyze perception performance under varying augmentation strengths. RISA is designed to generate physically plausible pseudo-samples that approximate real-world adverse conditions such as rain, snow, and fog, thereby improving generalization and robustness under environmental

degradation. This ablation focuses on two aspects of RISA's effectiveness: cross-modal physical consistency and alignment quality and stability across augmentation intensities (whether progressively stronger augmentations yield consistent performance gains).

In this experiment, we adopt a controlled "fixed-training and multi-level evaluation" protocol. Specifically, the model is trained with a fixed RISA-C3 augmentation setting, while both the SSL module and the VLM scheduler are disabled to isolate the standalone contribution of RISA. After training, the same model is evaluated on several held-out test sets generated with different RISA augmentation intensities from C1 to C5. The blank-control result is obtained by training the same model without RISA augmentation and evaluating it on the original held-out test set. Therefore, the C1-C5 rows in Table 3 reflect the robustness of the C3-trained model under different evaluation-time degradation intensities, rather than indicating that separate models are trained at each intensity level.

It should be emphasized that the fixed C3 setting in this subsection is introduced solely for controlled ablation purposes and does not

represent the training mode of the complete ROSE framework. In the full ROSE pipeline, degradation difficulty is organized into mixed scene - difficulty buckets, namely weather type $\times$ C1-C5, and the VLM scheduler dynamically adjusts the sampling probabilities across these buckets to support adaptive curriculum training. In addition to mAP, we further report the cross-modal re-projection error and BEV IoU under different evaluation intensities to quantify whether RISA preserves spatial consistency between image and LiDAR data. The experimental results are summarized in Table 3.

These results empirically demonstrate that RISA enhancement module helps alleviate detection performance attenuation caused by such interference to some degree. With the addition of low-to-moderate intensity samples, the model surpasses the original baseline without augmentation at all C1-C3 levels with a significant gain of mAP, with C3 performing the best. This quality can be attributed to two folds by examining its minimum re-projection error of 2.1 px and IoU of 99.1%: (1) such augmentations do not perturb either image or LiDAR sensors to the degree of causing modal misalignment; and (2) C3-level augmentation presents enough diversity of real-world corruptions such that the learned feature representation is enhanced instead of damaged. On the other hand, performance saturates in higher intensities C4 and C5. We hypothesize that the further increase of augmentation hurts detection results by introducing either modal misalignment or semantic corruption, which negatively affects the model's LiDAR-camera fusion learning. These degradations manifest themselves into higher re-projection error of 3.5 px and reduced IoU of 94.9% at C5 compared to the former. It is worth emphasizing that C3 being better than other levels of augmentations is not intended to imply an intrinsic superiority of such severity level of real roadside degradation - we do not claim that C1-C5 is a calibrated scale modeling real roadside weather degradation from light to severe. The perturbations designed in C1-C5 levels should only be viewed as controlled augmentation levels under a specific RISA-parameterized degradation family to analyze model robustness. Specifically, augmentations provided by current C1-C5 settings majorly take the form of "first-order" degradations: visibility drop, intensity attenuation and sparsification, and controllable pseudo-echo injection. Real roadside degradation may exhibit more complicated nonlinear effects such as spatially non-uniform raindrop lens blur, local occlusion, humidity-induced LiDAR multipath, and clustered spurious returns. Therefore, we view the saturating performance at C4/C5 as the penalty from overly strong perturbations imposed by our controlled degradation model instead of fully covering the severely corrupted real world.

Table 3. Impact of RISA augmentation intensity on perception performance

Evaluation intensity	Overall mAP (%)	E_{reproj} (px)	IoU (%)
ROSE-base	73.2	-	-
C1	74.0	3.2	95.6
C2	75.1	2.8	97.4
C3	75.6	2.1	99.1
C4	75.0	2.4	98.3
C5	74.2	3.5	94.9

To further disentangle the performance gain contributed by RISA from the whole ROSE pipeline, we further conducted a benchmark model training control experiment built upon RISA. Concretely, we re-trained several representative benchmark methods using the same RISA-C3 constructed physically plausible adverse weather enhanced training samples while keeping their network architectures, loss functions, and hyper-parameters totally unchanged. None of these benchmark models had SSL coupling or VLM-based curriculum

scheduling involved. We also ensured that all methods shared the same number of training iterations as well as the effective training sample budget. With this control experiment setting, we aimed to examine if the superiority of ROSE can still hold when the competing baselines were also benefited from the physically plausible adverse weather enhanced training.

All methods are evaluated on the same RISA-C3 augmented test set, while the difference lies in whether RISA-C3 augmentation is used during training. The experimental results are shown in Table 4. The results show that RISA-C3 training improves all baseline methods, especially in terms of robustness retention. This confirms that RISA is a transferable and effective augmentation module rather than a component that only benefits ROSE. However, the ROSE framework trained solely with RISA-C3 data still maintained certain advantages, especially in terms of the robustness performance corresponding to the RR metric. This result indicates that ROSE's final performance gain is not solely attributable to additional augmented samples, but also benefits from cross-modal SSL coupling and adaptive VLM-based curriculum scheduling. In other words, RISA improves the training domain coverage, while SSL coupling and VLM scheduling further enhance cross-modal consistency and difficulty-aware learning.

Table 4. Fair comparison with RISA-C3 augmented baseline training under the same evaluation protocol

Method	Training setting	mAP on RISA-C3 test (%)	RR (%)
BEVFusion	Original training	62.7	46.4
BEVFusion	RISA-C3 training	68.9	62.8
AutoAlignV2	Original training	63.4	76.7
AutoAlignV2	RISA-C3 training	70.2	82.6
ECFusion	Original training	64.8	50.7
ECFusion	RISA-C3 training	75.1	77.4
ROSE	RISA-C3 training	75.6	83.6

Overall, the RISA module plays a critical role in enhancing both fusion consistency and detection accuracy. It effectively expands the training domain with physically plausible and cross-modal consistent augmented samples, thereby improving the model's generalization under adverse conditions.

2) Effectiveness of the Self-Supervised Learning (SSL) Module

To assess the effectiveness of the self-supervised learning (SSL) coupling module in enhancing multi-modal perception, controlled experiments are conducted under consistent training settings—using C3-level RISA augmentation and no curriculum scheduling—to isolate the impact of different SSL configurations on detection accuracy (mAP) and fusion consistency (FCR). Four training setups are compared as follows:

- 1) baseline: Supervised learning only, with no self-supervised objectives.
- 2) ssl_cm: Cross-modal contrastive loss applied, enforcing semantic consistency between image and point cloud pillars in BEV space.
- 3) ssl_ts: Teacher-student consistency loss, where the student network is guided by stable teacher predictions under data perturbations.
- 4) ssl_both: A combined setup employing both cross-modal contrastive and teacher-student consistency losses.

All models share identical data, architectures, and optimization hyperparameters, with SSL losses being the only variable factor. The

final models were evaluated on a test set under natural weather conditions for both mAP and FCR. As shown in Figure 7, all configurations incorporating SSL accelerated convergence and achieved more stable training dynamics compared to the baseline. The ssl both setup yielded the smoothest loss curves and highest performance, indicating that jointly leveraging complementary self-supervised signals enhances representation learning, improves training efficiency, and contributes to greater stability.

In terms of detection accuracy, as shown in Figure 8, both self-supervised configurations—ssl ts and ssl cm—achieve noticeable improvements over the baseline model, which lacks any self-supervised signal. The cross-modal contrastive loss demonstrates a particularly strong effect on boosting detection performance. When both SSL signals are combined (ssl both), effectiveness of joint optimization.

As illustrated in Figure 9, the trend of Fusion Consistency Rate (FCR) closely mirrors that of mAP. Specifically, the cross-modal contrastive loss substantially improves FCR, indicating enhanced semantic alignment between image and point cloud modalities. Moreover, the ssl both configuration outperforms the individual components in FCR as well, validating the complementary nature of the two SSL objectives in enhancing fusion quality.

In summary, these results confirm that incorporating self-supervised learning mechanisms significantly improves both fusion consistency and detection accuracy in multi-modal perception tasks. While the contrastive loss is most effective for modality alignment, the teacher-student consistency loss enhances model robustness under perturbations. Their joint application achieves the best overall performance, validating the effectiveness and synergy of the proposed SSL coupling strategy.

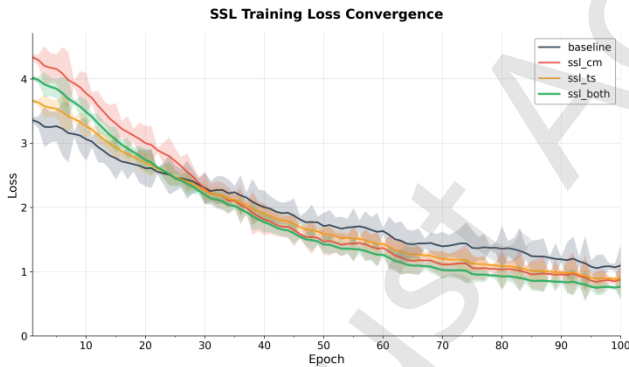


Fig. 7. Training dynamics under different SSL configurations: convergence speed and stability. Loss curves are reported for four training settings (baseline, ssl_cm, ssl_ts, ssl_both) under the same data/architecture. Introducing self-supervised signals accelerates convergence and stabilizes optimization, while combining cross-modal contrastive learning and teacher-student consistency (ssl_both) yields the smoothest and most stable training behavior.

3) Evaluation of the VLM scheduler Scheduling Mechanism

To further assess the role of the VLM scheduler in curriculum design, ablation studies are conducted with both RISA augmentation and SSL coupling enabled. Four curriculum scheduling strategies were compared:

- 1) Static Curriculum, a VLM-free control that follows a predefined and fixed difficulty progression;
- 2) Random Curriculum, a VLM-free control that stochastically samples difficulty levels without using VLM-generated semantic decisions;

3) Offline-VLM scheduling, in which curriculum strategies are generated by replaying historical training logs through the VLM scheduler;

4) Online-VLM scheduling, in which the curriculum strategy is updated online according to the current training feedback.

Static Curriculum and Random Curriculum serve as the two non-VLM scheduling controls in this ablation. All four strategies use the same training data, detector architecture, training budget, and evaluation protocol. Therefore, the experiment is designed to compare fixed and stochastic VLM-free scheduling with offline and online VLM-assisted scheduling.

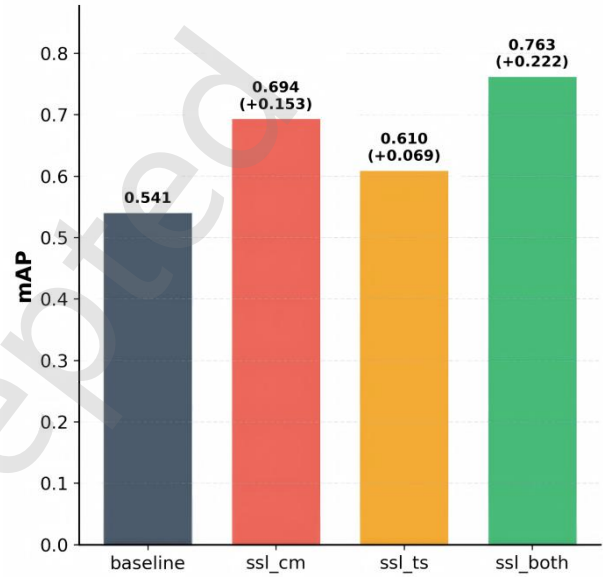


Fig. 8. Detection accuracy (mAP) comparison under different SSL objectives. This figure reports the final detection performance (mAP) for models trained with different SSL objectives. Both cross-modal contrastive learning (ssl_cm) and teacher-student consistency (ssl_ts) improve accuracy over supervised-only training, and their combination (ssl_both) achieves the best overall mAP, demonstrating complementary benefits.

All configurations used the same training data and model architecture, differing only in the curriculum scheduling mechanism. Therefore, the curriculum scheduling experiments evaluate a mixed-intensity training regime rather than a single fixed degradation level, which is more aligned with realistic deployment conditions where adverse-weather severity varies continuously across scenes. This setup allows for a quantitative evaluation of the marginal benefits introduced by the VLM scheduler.

The experimental results in Table 5 reveal clear performance differences across curriculum scheduling strategies in terms of detection accuracy (mAP), fusion consistency (FCR), robustness (R), and epoch efficiency measured by the number of epochs required to reach 95% of peak mAP. Overall, the random curriculum shows the lowest performance across all metrics, confirming that lack of structured guidance hinders effective model optimization. The static curriculum improves over random, but its rigid difficulty ordering lacks adaptability, limiting gains in both performance and data efficiency across training epochs.

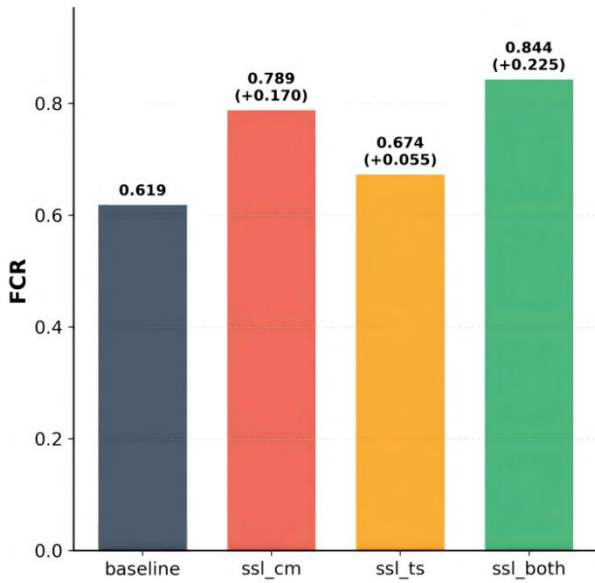


Fig. 9. Fusion consistency (FCR) comparison under different SSL objectives. Cross-modal contrastive learning significantly improves cross-modal feature alignment, and combining it with teacher-student consistency further enhances consistency, indicating better semantic correspondence between image and LiDAR representations.

By contrast, the VLM scheduler demonstrates clear advantages in curriculum design. The offline VLM scheduler outperforms both static and random strategies in mAP and FCR, while also achieving more stable robustness scores. Notably, the online VLM variant further surpasses its offline counterpart by dynamically adapting to semantic difficulty feedback during training. The online strategy attains the highest accuracy and consistency, the best retention rate in robustness, and the fewest epochs required to reach 95% of peak mAP, namely 68 epochs, compared with 86 and 94 epochs for the static and random strategies, respectively. The present experiments establish the effectiveness of the VLM-based realization adopted in ROSE, rather than the irreplaceability of VLMs for curriculum scheduling. Table 5 directly compares two VLM-free scheduling strategies, Static Curriculum and Random Curriculum, with Offline-VLM and Online-VLM scheduling. Static Curriculum follows a predetermined difficulty progression, whereas Random Curriculum changes the sampling schedule stochastically without using VLM semantic decisions. Both therefore provide direct controls for examining the effect of introducing VLM guidance. Under the same training and evaluation protocol, the Offline-VLM scheduler outperforms the Static and Random curricula in mAP, FCR, robustness retention, and convergence efficiency. Online-VLM further improves these metrics by using current training feedback to update the curriculum between scheduling rounds. Because the results are accumulated averages over multiple rounds under the declared random settings, the improvements persist across the scheduling trajectory rather than arising from a single favorable random outcome. These results demonstrate the stable benefit of the evaluated VLM-assisted realization relative to fixed and stochastic VLM-free scheduling. The purpose of this experiment is to isolate the effect of VLM participation under the evaluated curriculum settings, rather than to provide an exhaustive comparison with every possible handcrafted adaptive scheduling rule.

Table 5. Performance comparison of different curriculum strategies for VLM scheduler (The reported values are averages accumulated over multiple scheduling rounds under the stated random-seed settings.)

Strategy	Overall mAP	FCR	R	Epochs to 95% peak mAP
Static Curriculum	76.0	0.82	75.2	86
Random Curriculum	73.3	0.78	73.1	94
VLM offline	78.0	0.85	84.5	79
VLM online	78.8	0.86	89.6	68

The results validate that VLM scheduler can benefit multi-modal perception models. Specifically, the online version achieves a holistic superiority in accuracy, consistency, robustness as well as empirical epoch efficiency under identical training regime. This implies that the value of curriculum scheduling is not only from recommending suitable augmentation policy, but also releasing dynamically adjustable training signal. Moreover, it verifies that the VLM scheduler can provide helpful support and gains to multi-modal curriculum learning.

4.4 External Validation on Real-World Rainy and Foggy Roadside Data

To further examine the potential domain gap between RISA-synthesized degradations and real adverse weather conditions, we additionally conduct an external validation experiment using real-world rainy and foggy roadside data. Unlike the controllable robustness evaluation based on RISA-generated perturbations in the preceding sections, this experiment employs independently collected synchronized Camera-LiDAR data under real rainy and foggy conditions as a held-out external test set, aiming to assess the preliminary generalization capability of ROSE beyond the RISA-synthesized degradation domain.

The real adverse-weather dataset was collected from practical roadside perception scenarios and contains synchronized RGB images and LiDAR point clouds under measurable real-world rainy and foggy conditions. Representative examples are shown in Fig. 10. Vehicles, pedestrians, and cyclists in the dataset are manually annotated with both 2D and 3D bounding boxes. Specifically, the external test set consists of 7,973 frames of real rainy data and 1,980 frames of real foggy data, resulting in a total of 9,953 synchronized Camera-LiDAR frames and 41,228 annotated 3D objects. It should be emphasized that these real rainy and foggy data are used exclusively for final external testing. They are not involved in detector training, RISA parameter configuration or tuning, VLM scheduler curriculum generation, model selection, or hyperparameter adjustment. Therefore, this experiment is intended only to examine whether the detection performance learned from public datasets and RISA-augmented data can be partially transferred to real-world rainy and foggy roadside scenarios.

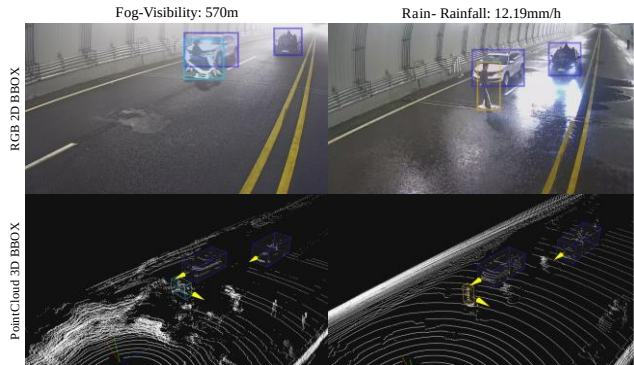


Fig. 10. Examples of Real-World Rainy and Foggy Data with Measured Weather Conditions.

To ensure fair comparison and clear interpretation, three

configurations are evaluated in the external validation experiment. The first configuration is ROSE-Base, namely the basic Camera-LiDAR fusion detector used in our framework. ROSE-Base adopts the same image encoder, LiDAR encoder, fusion neck, and 3D detection head as the full ROSE model, but disables RISA augmentation, SSL Coupling, and the VLM curriculum scheduler. It is trained only with the original public training set and the supervised detection loss. The second configuration is ROSE-Base +RISA, which introduces RISA-based synthetic augmentation into the same base detector while disabling SSL Coupling and the VLM scheduler. This configuration is used to investigate whether the physics-guided synthetic augmentation alone can provide transfer benefits under real rainy and foggy conditions. The third configuration is Full ROSE, which incorporates RISA, SSL Coupling, and the VLM scheduler, and is used to evaluate the performance of the complete closed-loop enhancement framework on the real adverse-weather external test set. All models follow the same training protocol as in the main experiments and are directly evaluated on the real rainy and foggy test set. Following the 3D detection protocol used in the preceding sections, we report the mAP on the rainy subset, the foggy subset, and the overall test set. The experimental results are presented in Table 6.

Table 6. External validation on real-world rainy/foggy roadside data.

Method	Training setting	Rain mAP	Fog mAP	Overall mAP
ROSE-Base	Public training data only	61.8	55.4	58.7
ROSE-Base +RISA	Public training data+RISA	65.2	59.8	62.6
Full ROSE	Public training data+RISA+SSL+VLM	68.7	63.5	66.2

The results show that ROSE-Base+RISA achieves higher detection accuracy than ROSE-Base on the real rainy and foggy data. This indicates that the physics-guided cross-modal synthetic augmentation is not only effective within the controllable perturbation domain defined by RISA, but can also improve the model's adaptability to real rainy and foggy degradations to a certain extent. Building upon this, Full ROSE further achieves the best performance. Its overall mAP improves by 7.5 points over ROSE-Base and by 3.6 points over ROSE-Base+RISA. On the rainy and foggy subsets, Full ROSE achieves mAP values of 68.7 and 63.5, respectively, outperforming the compared configurations in both cases. These results suggest that the cross-modal consistency learning introduced by SSL Coupling and the feedback-driven curriculum scheduling enabled by the VLM scheduler can further enhance the model's generalization ability under adverse weather conditions.

Note that this external validation experiment is not meant to suggest that the synthetic-to-real gap has been entirely closed. Actual rainy/foggy conditions may include complex combinations of degradation not yet well represented by the current RISA model formulation. Therefore, this experiment is to be viewed as external validation with small-scale real-world data, rather than thorough validation on a comprehensive real adverse-weather benchmark. However, these results do demonstrate initial evidence that ROSE's robustness improvements do not merely apply to synthetic perturbations produced by RISA, but can transfer to independently collected real-world rainy/foggy roadside conditions as well. Multi-location validation with larger scale/variation adverse-weather datasets will be performed in future work.

4.5 Discussion

While the ablation results validate the effectiveness of each major component of ROSE, including RISA, SSL coupling and VLM scheduler, we also discuss more about the implications and limitations of multi-

modal perception in our work. Indeed, depending on the scenarios, even adversarial weather conditions, it might still be possible to achieve competitive performance using single-modality model that solely relies on either camera images or LiDAR. Single-modality perception is limited by the failure modes of the corresponding sensor. Failures for cameras can happen due to low illumination or glare as well as visibility degradation in rain/fog/snow. Failures for LiDAR can also happen due to illumination changes due to its intensity signals, but it is still prone to attenuation, sparsification and spurious returns when the environment is degraded. Multi-modal perception complements each other in terms of semantic and geometric cues. When one fails, the other could likely overcome its shortcomings. Thus, instead of thinking of multi-modal perception as just a way to increase detection performance, it should also be viewed as a way to increase robustness, failure tolerance as well as semantic completeness under complex disturbances.

In other words, ROSE does not aim to show a generalized superiority of all multi-modal systems over all single-modality pipelines under all possible conditions. Instead, we explore if and how physically inspired augmentation and cross-modal consistency learning lead to better perception stability under bad weather. The distinction is critical because ROSE's primary contribution is enhanced multi-modal collaboration with aligned degradation simulation and consistency-driven representation learning, not modality comparison. This goal is especially significant for roadside perception where long-term stability and robustness are valued more than short-term peak metrics.

At the same time, several limitations of the current study should be acknowledged. First, the present framework is mainly developed for Camera-LiDAR roadside multi-modal perception, and the current RISA module does not yet fully capture more complex sensor degradation mechanisms, such as raindrop-induced lens blur, highly nonlinear camera response, or LiDAR multipath effects in high-humidity environments. Second, the current experimental validation is still limited to the adopted benchmark setting, and broader cross-dataset and real-world evaluation is needed to further assess generalization. Third, although the present study focuses on multi-modal consistency rather than single-modality comparison, more explicit single-modality and heterogeneous-modality analyses would be valuable in future work for better quantifying the independent contribution of multi-modal collaboration. Fourth, the current scheduler ablation does not include a non-VLM adaptive heuristic curriculum using the same feedback signals.

Despite these limitations, the proposed framework shows promising practical potential in adverse-weather roadside perception enhancement, resilient V2X cooperative perception, all-weather traffic monitoring, and low-annotation-cost expansion of multi-modal training data. Future work will further incorporate more comprehensive single-modality comparison experiments, finer-grained physical degradation modeling, and closer co-optimization between roadside and onboard perception systems.

5 Conclusions

This paper introduces ROSE, a unified framework for training roadside multimodal perception models under challenging weather conditions, where limited adverse-weather data, degraded cross-modal semantics, and unstable optimization often restrict model performance. ROSE consists of three main components: the first-order physics-guided cross-modal augmentation module (RISA), a self-supervised cross-modal coupling strategy, and a VLM-based curriculum scheduler. Rather than operating independently, these components jointly regulate data augmentation, representation learning, and the progression of training difficulty. Extensive experiments show that ROSE consistently improves detection accuracy, fusion consistency, and robustness over the evaluated baselines. In particular, samples generated by RISA provide clear benefits under relatively rare and long-tail conditions,

including rain, snow, and fog. The VLM scheduler further adjusts the training curriculum according to feedback from the current learning state, making limited bucket-level updates to curriculum difficulty instead of relying on a fixed sampling strategy throughout training. In this sense, ROSE replaces a fully static training schedule with a data-dependent curriculum that can adapt as training proceeds. The VLM itself remains outside the detector optimization loop and is used only as a weak scheduler; it does not directly update or optimize the detector parameters. The current RISA degradation space should nevertheless be viewed as a first-order approximation of real roadside sensor degradation rather than a complete reproduction of all nonlinear effects encountered in practice, such as sensor saturation. Further improvements will require calibrated real-world degradation data and more detailed sensor-specific models, for example, spatially non-uniform image blur caused by raindrops or humidity-related multipath effects in LiDAR measurements. Another useful direction is to define sim-to-real degradation-gap metrics that quantify the discrepancy between augmented and real adverse-weather data and feed this information back into curriculum scheduling. These extensions could improve the physical fidelity of adverse-weather augmentation and make ROSE more applicable to real roadside perception deployments.

Author contributions

Guoyu Zhang: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Peng Hang: Writing – review & editing, Supervision, Methodology, Funding acquisition.

Xin Xia: Review & editing, Formal Analysis, Methodology.

Jian Sun: Review & editing, Formal Analysis, Methodology.

Replication and data sharing

For reproducibility and verification, we have organized and open-sourced the main algorithmic implementation and experimental pipeline used in this study. The codebase, configurations, and scripts can be found at: [DOI: 10.26599/ETSD.2026.9190081](https://doi.org/10.26599/ETSD.2026.9190081).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (52302502, 52472451), the Shanghai Scientific Innovation Foundation (No.23DZ1203400), and the Fundamental Research Funds for the Central Universities.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language polishing, grammatical correction, and improving the clarity and readability of the manuscript. The tool was not used to generate research data, experimental results, or scientific conclusions. After using this tool, the authors thoroughly reviewed, verified, and edited all relevant content as necessary and take full responsibility for the final content of the published article.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Afham, M., Nuha, I.T., Nguyen, H.D., Ferdous, S.M.M., Vu, N.A., 2022. CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D

Point Cloud Understanding, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9892–9902.

Aloufi, N., Alnori, A., Basuhail, A., 2024. Enhancing Autonomous Vehicle Perception in Adverse Weather: A Multi Objectives Model for Integrated Weather Classification and Object Detection. *Electronics* 13(15), 3063.

Asif-Ur-Rahman, M., Afsana, F., Mahmud, M., Kaiser, M.S., Ahmed, M.R., Kaiwartya, O. et al., 2019. Toward a Heterogeneous Mist, Fog, and Cloud-Based Framework for the Internet of Healthcare Things. *IEEE Internet of Things Journal* 6(3), 4049–4062.

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H. et al., 2022. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection With Transformers, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1090–1099.

Bernuth, A.v., Volk, G., Bringmann, O., 2019. Simulating Photo-realistic Snow and Fog on Existing Images for Enhanced CNN Training and Evaluation, In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 41–46.

Bijelic, M., Gruber, T., Mannan, F., Dietmayer, K., Heide, F., 2020. Seeing Through Fog Without Seeing Fog: Deep multi-modal Sensor Fusion in Unseen Adverse Weather, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11682–11692.

Cai, M., Luo, C., Lee, Y.J., Yang, X., 2024. Cross-Modal Self-Supervised Learning with Effective Contrastive Units for LiDAR Point Clouds, In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 9468–9475.

Chen, J., Zhang, M., Liu, Q., 2023. Semantic Scene Understanding with Depth Information for Robust Perception in Foggy Environments. *IEEE Trans Veh Technol* 72(5), 1234–1245.

Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F., 2022. Deformable Feature Aggregation for Dynamic Multi-modal 3D Object Detection, In: European conference on computer vision, 628–644.

Cho, M., Cao, Y., Sun, J., Zhang, Q., Pavone, M., Park, J.J. et al., 2025. Cocoon: Robust Multi-Modal Perception with Uncertainty-Aware Sensor Fusion. *International Conference on Learning Representations, ICLR 2025*.

Choi, S., Kim, J., Shin, H., Choi, J.W., 2025. Mask2Map: Vectorized HD Map Construction Using Bird's Eye View Segmentation Masks, *European Conference on Computer Vision*, 19–36.

Choo, Y.-S., Kim, B., Kim, H.-S., Park, Y.-S., 2024. Supervised Contrastive Learning for 3D Cross-Modal Retrieval. *Applied Sciences* 14(22), 10322.

Duy-Tho Le, H.S., Jianfei Cai, Hamid Rezatofighi, 2025. Diffusion Model for Robust Multi-sensor Fusion in 3D Object Detection and BEV Segmentation, In *European Conference on Computer Vision*, 232–249.

El-Shair, Z.A., Hossain, M.M., Islam, M.S., 2024. SID: Stereo Image Dataset for Autonomous Driving in Adverse Conditions, In: NAECON 2024-IEEE National Aerospace and Electronics Conference, IEEE, 403–408.

Fang, S., Zhao, X., Liu, X., Hang, P., Wang, J., Wang, Y. et al., 2026. Toward Cooperative Driving in Mixed-Traffic: An Adaptive Potential Game-Based Approach with Field Test Verification. *Engineering*. <https://doi.org/10.1016/j.eng.2026.03.022>

Fu, J., Gao, C., Wang, Z., Yang, L., Wang, X., Mu, B. et al., 2024. Eliminating Cross-modal Conflicts in BEV Space for LiDAR-Camera 3D Object Detection, In: 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 6381–16387.

Gao, Z., Jia, B., Xie, D., Wang, W. and Wu, J., 2025. A Discussion on the Complexity and Transit Mechanisms of Urban Traffic Systems. *Engineering* 44(1), 26 – 31.

Guo, Q., Liang, J., Wang, H., 2023. Night Vision Anti-Halation Algorithm of Different-Source Image Fusion Based on Low-Frequency Sequence Generation. *Mathematics* 11(10), 2237.

Hahner, M., Sakaridis, C., Dai, D., Gool, L.V., 2022. LiDAR Snowfall Simulation for Robust 3D Object Detection, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 16364–16374.

Haider, A., Pigniczki, M., Koyama, S., Köhler, M.H., Haas, L., Fink, M. et al.,

2023. A Methodology to Model the Rain and Fog Effect on the Performance of Automotive LiDAR Sensors. *Sensors* 23(15), 6891.
- Halder, S.S., Lalonde, J.-F., Charette, R.d., 2019. Physics-Based Rendering for Improving Robustness to Rain, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10203-10212.
- Hao R Y,Fan S Q,Dai Y R,et al., 2024. RCooper:a real-world large-scale dataset for roadside cooperative perception In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,Seattle,WA,USA,2024:22347-22357.
- Heinzler, R., Piewak, F., Schindler, P., Stork, W., 2020. CNN-Based Lidar Point Cloud De-Noising in Adverse Weather. *IEEE Robotics and Automation Letters* 5(2), 2514-2521.
- Kim, J., Park, B.-j., Kim, J., 2023. Empirical Analysis of Autonomous Vehicle's LiDAR Detection Performance Degradation for Actual Road Driving in Rain and Fog. *Sensors* 23(6), 2972.
- Kumar, D., Muhammad, N., 2023. Object Detection in Adverse Weather for Autonomous Driving through Data Merging and YOLOv8. *Sensors* 23(20), 8471.
- Li, B., Liu, P., Fu, L., Li, J., Fang, J., Xu, Z., Yu, H., 2025. Enhancing vehicle re-identification by pair- flexible pose guided vehicle image synthesis. *Green Energy and Intelligent Transportation* 4(5), 100269.
- Li, K., Cui, H., Dai, Z., Song, H., 2024. RVIFNet: Radar-Visual Information Fusion Network for All-Weather Vehicle Perception in Roadside Monitoring Scenarios. *IEEE Sensors Journal* 24(23), 40105-40122.
- Li, X., Hu, Z., Xiao, X. and Kong, B., 2026. Look Before You Fuse: 2D-Guided Cross-Modal Alignment for Robust 3D Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11589 - 11598.
- Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D. et al., 2022. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17182-17191.
- Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y. et al., 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, 2774-2781.
- Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y. et al., 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems* 35, 10421-10434.
- Liu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., & Qu, X. (2023). Can language models be used for real-world urban-delivery route optimization?. *The Innovation*, 4(6), 100520.
- Liu, Y., Zhou, Y., Liu, Y., Xu, Z., & He, Y. (2025). Intelligent fault diagnosis for CNC through the integration of large language models and domain knowledge graphs. *Engineering*, 53(10), 311 - 322.
- Lu, N., Cheng, N., Zhang, N., Shen, X., Mark, J.W., 2014. Connected Vehicles: Solutions and Challenges. *IEEE Internet of Things Journal* 1(4), 289-299.
- Lu, Q., Wang, X., Jiang, Y. et al., 2025. OmniTester: Multimodal Large Language Model Driven Scenario Testing for Autonomous Vehicles. *Automotive Innovation* 8, 838 - 852.
- M. Hahner, C.S., D. Dai, and L. Van Gool, 2021. Fog Simulation on Real LiDAR Point Clouds for 3D Object Detection in Adverse Weather, In: *Proceedings of the IEEE/CVF international conference on computer vision*, 15283-15293.
- Massoud, Y., Laganieri, R., 2024. Learnable fusion mechanisms for multi-modal object detection in autonomous vehicles. *IET Computer Vision* 18(4), 499-511.
- Musat, V., Fursa, I., Newman, P., Cuzzolin, F., Bradley, A., 2021. Multi-weather city: Adverse weather stacking for autonomous driving, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2906-2915.
- Qian, H., Wang, M., Zhu, M., Wang, H., 2025. A Review of Multi-Sensor Fusion in Autonomous Driving. *Sensors* 25(19), 6033.
- Qin, Y., Wang, C., Kang, Z., Ma, N., Zhang, R., 2023. SupFusion: Supervised LiDAR-Camera Fusion for 3D Object Detection, In: *Proceedings of the IEEE/CVF international conference on computer vision*, 22014-22024.
- S. Teufel, G.V., S. Gamberdinger, and O. Bringmann, 2023. Enhancing Robustness of LiDAR-Based Perception in Adverse Weather using Point Cloud Augmentations, In: *2023 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 1-6.
- Sakaridis, C., Wang, H., Li, K., Zurbrugg, R., Jadon, A., Abbeloos, W. et al., 2021. ACDC: The Adverse Conditions Dataset With Correspondences for Semantic Driving Scene Understanding, In: *Proceedings of the IEEE/CVF international conference on computer vision*, 10765-10775.
- Sezgin, F., Steinhauser, D., Vriesman, D., Lugner, R., Brandmeier, T., 2023. Safe autonomous driving in adverse weather: Sensor performance evaluation and degradation detection. *IEEE Trans Intell Veh* 8(1), 1-12.
- Sheeny, M., Hossain, M.M., Islam, M.S., 2020. RADIATE: A Radar Dataset for Automotive Perception in Bad Weather. *IEEE Trans Intell Veh* 5(3), 442-453.
- Shi, K., Feng, Y., Hu, T., Cao, Y., Wu, P., Liang, Y. et al., 2025. FusionNet: Multi-model Linear Fusion Framework for Low-light Image Enhancement, In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1028-1037.
- Teufel, S., Volk, G., Bernuth, A.v., Bringmann, O., 2022. Simulating Realistic Rain, Snow, and Fog Variations for Comprehensive Performance Characterization of LiDAR Perception, In: *2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring)*, IEEE, 1-7.
- Vilsan, A., Sandu, C., Anghelache, G., Warfford, J., 2025. Development of an intelligent instrumented tire for real- time hydroplaning risk estimation. *SAE International Journal of Vehicle Dynamics, Stability, and NVH* 10(3), 439-454.
- Vinoth, K., Sasikumar, P., 2024. Multi-sensor fusion and segmentation for autonomous vehicle multi-object tracking using deep Q networks. *Scientific Reports* 14(1), 31130.
- Vora, S., Lang, A.H., Helou, B., Beijbom, O., 2020. PointPainting: Sequential Fusion for 3D Object Detection, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4604-4612.
- Wang, C., Ma, C., Zhu, M., Yang, X., 2021. PointAugmenting: Cross Modal Augmentation for 3D Object Detection, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11794-1180.
- Wang, J., Wu, Z., Liang, Y., Tang, J., Chen, H., 2024a. Perception Methods for Adverse Weather Based on Vehicle Infrastructure Cooperation System: A Review. *Sensors* 24(2), 374.
- Wang, L., Li, Y., Liu, H., 2023. Depth-Aware Weather Removal for Semantic Segmentation in Adverse Weather Conditions. *IEEE Trans Image Process* 32, 4567-4579.
- Wang, X., Tian, J., Yu, Y., Nyengor Agbenu, I.K., Wang, Q., Feng, Y. et al., 2024b. LIASM-NRID: Constructing an atmospheric scattering model for low-light conditions and dehazing nighttime road images. *Optics Communications* 569, 130803.
- Wei, K., Fu, Y., Yang, J., Huang, H., 2020. A Physics-based Noise Formation Model for Extreme Low-light Raw Denoising, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2758-2767.
- Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F. et al., 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher, In: *Proceedings of the IEEE/CVF international conference on computer vision*, 3060-3069.
- Yang, J., Shi, S., Wang, Z., Li, H., Qi, X., 2021. ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10368-10378.
- Yin, J., Shen, J., Chen, R., Li, W., Yang, R., Frossard, P. et al., 2024. IS Fusion: Instance-Scene Collaborative Fusion for multi-modal 3D Object Detection, In: *Proceedings of the IEEE/CVF conference on computer vision and*

pattern recognition, 14905-14915.

- You, J., Gan, R., Tang, W., Huang, Z., Liu, J., Jiang, Z. et al., 2025. FollowGen: A scaled noise conditional diffusion model for car-following trajectory prediction. *Communications in Transportation Research* 5, 100215.
- Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y. et al., 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection, In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21361-21370.
- Zhang, J., Zhan, R., Wang, Y., Qu, X., 2025. Optical communication based V2V for vehicle platooning. *Green Energy and Intelligent Transportation* 4(4), 100278.
- Zang, S., Ding, M., Smith, D., Tyler, P., Rakotoarivelo, T., Kaafar, M.A., 2019. The Impact of Adverse Weather Conditions on Autonomous Vehicles: How Rain, Snow, Fog, and Hail Affect the Performance of a Self-Driving Car. *IEEE Vehicular Technology Magazine* 14(2), 103-111.
- Zhang, Y., Carballo, A., Yang, H., Takeda, K., 2023. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing* 196, 146-177.
- Zhao, X., Wen, C., Zhu, X., Wang, Y., Bai, H. and Dou, W., 2025. TripleMixer: A Triple-Domain Mixing Model for Point Cloud Denoising Under Adverse Weather. *IEEE Transactions on Image Processing* 34, 7712 - 7727.

Author biography (mandatory)



Guoyu Zhang received the M.S. degree in Electrical Engineering from Curtin University, Perth, Australia, in 2020. He is currently pursuing the Ph.D. degree with the School of Transportation Engineering, Tongji University, Shanghai, China. His current research interests include roadside multi-modal perception optimization and intelligent transportation systems.



Peng Hang is an Associate Professor at the Department of Traffic Engineering, Tongji University, Shanghai, China. He received the Ph.D. degree with the School of Automotive Studies, Tongji University, Shanghai, China, in 2019. He was a Visiting Researcher with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, in 2018. From 2020 to 2022, he served as a Research Fellow with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His research interests include vehicle dynamics and control, decision making, motion planning and motion control for autonomous vehicles. He serves as an Associate Editor of *IEEE Internet of Things Journal*, *IEEE Transactions on Vehicular Technology*, *Journal of Field Robotics*, *IET Smart Cities*, and *SAE International Journal of Vehicle Dynamics, Stability, and NVH*.



Xin Xia was an Assistant Professional Researcher at the Department of Civil and Environmental Engineering, University of California at Los Angeles, Los Angeles, CA, USA. He also worked as a Post-Doctoral Fellow with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently an Assistant Professor at the Department of Mechanical Engineering, University of Michigan-Dearborn, Dearborn, MI, USA. His research interests include automated driving and cooperative driving automation with a focus on cooperative perception, multisensor fusion state estimation, localization and mapping, and smart infrastructure. Dr. Xia received the award from the Intersection Safety System Challenge from USDOT and was a best paper award finalist at the 36th FISITA World Automotive Congress. He serves as an Associate Editor for *IEEE Transactions on Intelligent Vehicles*, *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, and *IET Intelligent Transportation Systems*. He also serves as the Paper Review Coordinator for Transportation Research Board ACP30 and the Award and Challenge Chair of IEEE ITSC 2024.



Jian Sun received the Ph.D. degree from Tongji University in 2006. Subsequently, he was at Tongji University as a Lecturer, and then promoted to the position as a Professor in 2011, where he is currently a Professor with the College of Transportation Engineering and the Dean of the Department of Traffic Engineering. His main research interests include traffic flow theory, traffic simulation, connected vehicle/infrastructure system, and intelligent transportation systems.