Multimodal Machine Learning-Based Unmanned Driving System

Jiayang Liang*, Shuo Zhao
Electronic and Automation College, City Institute, Dalian University of Technology, Dalian, Liaoning, China
**Corresponding Author*

Abstract: With the breakthrough of artificial intelligence technology, unmanned driving system are moving from theory to practical application. However, they face multiple disturbances such as sensor anomalies and extreme weather in real dynamic and complex environments, posing severe challenges to the safe and reliable operation of the systems. Although machine learning technology has significantly improved system performance, it still has insufficient robustness when dealing with these challenges. Therefore, in-depth research and improvement of system robustness are crucial for ensuring driving safety and facilitating the application of the technology. This paper focuses on machine learning-based unmanned driving systems and conducts a systematic study of their robustness issues. The study begins by sorting out the core robustness challenges that the system faces at the levels of perception, decision-making, and control. Based on this, this paper explores comprehensive robustness enhancement methods including data augmentation, adversarial training, regularization, redundant design, and high-fidelity simulation testing from two dimensions: internal enhancement of machine learning algorithms and external assurance of system architecture. The findings suggest that enhancing the robustness of unmanned driving system is a systematic project that requires a combination of algorithm optimization, architectural fault tolerance, and complete verification. Significant progress has been made through the combination of multiple levels of technology at present, but in the future, there is a need to further evolve from "robustness" in response to known threats to "system resilience" with adaptive capabilities. Finally, closing the verification gap between simulation and reality and achieving collaborative optimization of multiple technical paths are key directions for building truly safe and reliable unmanned driving system.

Keywords: Machine Learning; Driverless Technology; Robustness Research

1. Introduction

The rapid development of artificial intelligence technology is pushing unmanned driving system from theoretical research to practical application. At its core, it relies on the coordinated operation of three modules: environmental perception, decision-making planning, and control execution, to enable autonomous operation of the vehicle. However, the real traffic environment is full of dynamic complexity and high uncertainty, which poses a serious challenge to the robustness of the system-that is, its ability to operate stably and reliably under non-ideal conditions such as noise interference, sensor anomalies, and extreme weather. Robustness has become a key attribute for ensuring the safety of driverless vehicles and achieving large-scale application of the technology.

Currently, deep learning technology has achieved great success in the field of unmanned driving, with the advantages of high accuracy, strong robustness, and low[1]. Although machine learning methods such as deep learning have significantly improved the accuracy of perception and decision-making, the performance of systems can still significantly deteriorate when facing partial sensor failure, bad weather, or rare scenarios other than training data. Decision-making modules also often lag in response to unexpected situations due to insufficient generalization ability of the model, and the lack of explainability caused by their "black box" nature further raises questions about technical credibility and accountability. Therefore, conducting research on the robustness of unmanned driving system is not only an inherent need for the deepening development of the technology, but also a fundamental basis for

promoting its commercial application and building social trust.

In response to these issues, this study focuses on machine learning-based unmanned driving system, aiming to systematically analyze their weak points in robustness in complex dynamic environments and explore practical paths to enhance them. First, it is necessary to clarify the main types of disturbances that the system faces in the actual environment, such as sensor noise, abnormal input data, and non-steady-state changes in environmental distribution, and analyze the specific impact mechanisms of these disturbances on core modules such as perception, positioning, and decisionmaking, so as to systematically identify the key nodes for robustness improvement. Secondly, a multi-dimensional quantitative evaluation system for robustness must be constructed. This requires a hierarchical modeling of actual road scenarios, particularly covering extreme weather, sudden obstacles, sensor failures and other extreme conditions, and quantifying the performance degradation patterns of the system under different types and intensities of disturbances through a

combination of simulation tests and real road tests, and then establishing a correlation model between disturbance intensity and system failure probability.

For the system vulnerabilities revealed by the assessment, a hierarchical reinforcement strategy is needed to reinforce them. At the perceptual level, fault tolerance can be enhanced through multimodal data fusion and redundant verification mechanisms, and neural network resistance to noise and adversarial samples can be strengthened by methods such as adversarial training. At the decision-making level, dynamic fault tolerance and adaptive strategies based on reinforcement learning can be introduced to enable the system to adapt online to changes in data distribution. At the control level, a multi-level security degradation mechanism should be designed to automatically switch to the basic security mode when an anomaly is detected. These enhancements need to be tailored closely to the computational characteristics and real-time requirements of each module.

Technically, this study follows a closed-loop framework of "test-hardening-verification". The test phase will build a benchmark test set that includes multiple types of interference factors such as light variations, occlusions, and sensor failures; The hardening phase, based on the modular design principle, integrates adaptive robustness enhancement techniques for each subsystem; The validation phase will use theoretical tools, such as Koopman operator theory, to analyze the stability boundary of the system under disturbances, ultimately forming a set of full life cycle robustness assurance solutions that cover international standard requirements. Through this systematic study, the aim is to enhance the adaptability and safety redundancy of the unmanned driving system in complex open environments by comprehensively applying algorithmic optimization, multimodal fusion, explainability enhancement and safety architecture design, thereby providing solid technical support for its reliable deployment and wide application.

2. Current Situation and Problems of Unmanned Driving Systems

2.1 Current Status of Unmanned Driving System

Driverless technology is a cutting-edge technology that relies on computers and artificial intelligence to complete, safe and efficient driving without human manipulation. In the 21st century, due to the increasing number of car users, road traffic is facing more and more serious problems such as congestion and frequent safety accidents. With the support of Internet of Vehicles technology and artificial intelligence technology, driverless technology can coordinate travel routes and planning time, thereby significantly improving travel efficiency and reducing energy consumption to a certain extent. Driverless cars can also reduce safety hazards such as drunk driving and fatigued driving, reduce driver errors and enhance safety. Driverless technology has thus become a key focus of research and development around the world in recent years. Unmanned driving system are a complex system. In order to achieve the driving process from point A to point B, in the actual use of driverless vehicles, the driverless system needs to complete three major tasks: perception, decision-making, and control. The three modules are the environmental perception system, the positioning and navigation system, and the path planning system, as shown in Figure 1. The primary condition for achieving unmanned driving is to perceive the surrounding environment of the vehicle through "seeing" and

"hearing". The perception system relies on large amounts of data from sensors to perceive and monitor the vehicle's movement, the environment, and the driver's state and behavior. The unmanned driving perception system uses a wide variety of sensors, including cameras, millimeter-wave radars, lidars, ultrasonic radars, infrared night vision, as well as GNSS(Global Navigation Satellite System) and IMU(Inertial Measurement Unit) for positioning and navigation. Another type of technology, though not an active detection element, is a collaborative global data assistance that can expand the environmental perception capabilities of intelligent vehicles and also plays an indispensable role in the perception system. These technologies include high-precision maps, V2X vehicle-to-everything systems, etc. Each type of perception technology has its own advantages and disadvantages, and they are fully integrated with each other to form comprehensive and reliable perception data for decision-making and control systems[2].

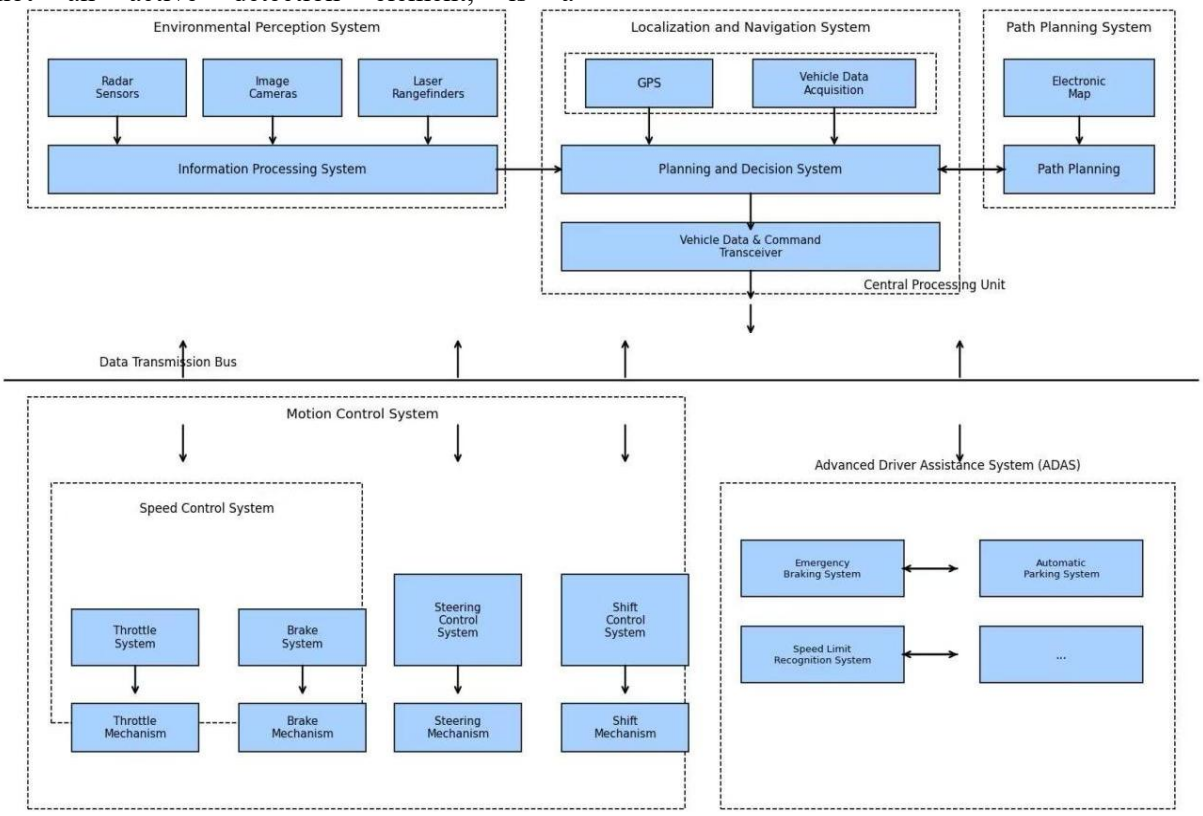

Figure 1. Technical Architecture of Unmanned Driving System

2.2 Analysis of Robustness Issues of Unmanned Driving System

The term "Robustness" is derived from the fields of control engineering and computer science and refers to the ability of a system to exhibit stability and reliability under conditions of uncertainty. In simple terms, robustness describes the ability of a system to maintain normal functioning in the face of external disturbances, unexpected situations, or extreme conditions.

In the field of unmanned driving, robustness is mainly reflected in scenarios where a vehicle can still operate safely and stably in the face of uncertainties such as complex road conditions, weather changes, and sensor failures. An autonomous vehicle may show good driving performance when traveling on a clear highway. But whether a system can continue to safely complete driving tasks in the event of sudden heavy rain, heavy snow, blurred cameras due to dirt, or lost GPS signals depends on the level of its robustness. The complexity of unmanned driving lies in the fact that it needs to combine the robustness of multiple links, including perception, decision-making, control and communication. These modules not only need to be robust separately, but also be able to work in coordination to ensure overall stability through compensation mechanisms when problems occur locally.

The issue of robustness in unmanned driving system involves the entire process of perception,

decisionmaking, and control. Based on the research progress in 2023-2024, it can be classified into the following typical problems[3]:

(1) Problems with the robustness of the perception system

Vulnerability to adversarial attacks: The system is vulnerable to both digital and physical adversarial attacks. For example, minor perturbations in an image can lead to misidentification of signs (such as YOLOv5 recognizing a stop sign as a speed limit sign); Deceiving LiDAR with an infrared laser or sticking a specific pattern on a road sign can also interfere with the perception results.

Environmental interference sensitivity: Extreme weather conditions such as heavy rain and heavy snow can cause camera images to blur and LiDAR point clouds to decay, significantly reducing detection performance. In addition, fusion algorithms may fail when data from multiple sensors (such as cameras and radars) are inconsistent under anomalous conditions.

(2) Robustness issues of the decision system

Failure in long-tail scenarios: When faced with rare scenarios in the training data, such as temporary construction signs, sudden animal intrusions, the decision model is prone to mishandling, resulting in planning failure or the need for frequent human takeover.

Multi-agent interaction uncertainty: In complex interactions with human drivers, pedestrians, etc., the model has difficulty accurately predicting the intentions of the other party (such as ramp merge game, pedestrian hesitation behavior), thereby increasing the risk of collision.

(3) Control system robustness issues

Actuator dynamic response limitations: In uncertain dynamic conditions such as lowadhesion road surfaces (such as ice), control algorithms (such as MPC) may not be able to precisely track trajectories beyond safety limits.

Hardware failure propagation: The failure of a single sensor (such as a camera) may trigger a chain reaction, leading to a degradation in perception performance, which in turn forces the planner to adopt an overly conservative strategy, affecting the overall system efficiency.

(4) Issues with data and training robustness

Simulation-reality gap: Models trained in simulated environments such as CARLA often experience significant performance degradation when transferred to the real world due to differences in lighting, textures, etc.

Annotation noise amplification: Annotation errors in the training data, such as incorrect 3D bounding boxes, are learned and amplified by the model, resulting in a higher false detection rate in real-world applications.

(5) System-level robustness issues

Module coupling failure: The subsystems are tightly coupled, and anomalies in a single module (such as sensing delay) propagate along the link, resulting in system-level performance degradation such as planning lag and control oscillation.

Cyber security threats: If the communication link between the vehicle and the external environment (V2X) is hijacked, forged information (such as false traffic light status) may trigger dangerous group behavior (such as sudden braking by a convoy).

These issues indicate that the robustness challenges of unmanned driving system span across all levels from data and algorithms to hardware and systems, and are key bottlenecks restricting their safe and reliable deployment.

2.3 Chapter Summary

This chapter elaborates on the current state of unmanned driving system and the robustness challenges they face. First, an analysis of the basic architecture for autonomous operation of unmanned driving system through the three core modules of perception, decision-making, and control is presented, and the progress made in current technologies in areas such as multi-sensor fusion, high-precision positioning, and intelligent planning is outlined. These advancements have laid the foundation for driverless technology to move from concept to practical application.

However, the complexity of the system also brings about multi-dimensional robustness problems. This chapter systematically reviews the main challenges facing current technologies at five levels: perception, decision-making, control, data training, and the system as a whole: At the perception level, the system is vulnerable to adversarial attacks and environmental disturbances; At the decision-making level, models struggle to handle long-tail scenarios and uncertain multi-agent interactions; At the control level, there is a risk of actuator dynamic response limitations and hardware failure propagation; At the data and training level, the gap between simulation and reality, as well as annotation noise, can significantly affect model performance; At the system level, tight coupling

between modules and cybersecurity threats pose additional systemic risks. These challenges are interrelated and span the entire technology stack from underlying data to top-level systems, collectively forming the core bottlenecks that unmanned driving system must overcome to move towards large-scale, high-security deployments. A clear definition and in-depth analysis of these issues is a prerequisite for subsequent research to propose targeted robustness enhancement solutions.

3. Applications and Challenges of Machine Learning in Unmanned driving system

Machine learning is an important aspect of artificial intelligence. In plain terms, it is about giving machines the ability to learn[4]. At present, machine learning, especially deep learning models, has become a key technology driving the development of core links such as perception, decision-making, planning and control in unmanned driving system. This chapter aims to explore the current application status of various machine learning models in the field of driverless vehicles and analyze the key technical paths adopted to enhance system robustness, including data processing in the input layer, regularization mechanisms in the hidden layer, and loss optimization in the output layer.

3.1 Overview of the Application of Machine Learning Models in Driverless Vehicles

The unmanned driving system widely integrates a variety of cutting-edge machine learning models. Large language models (LLMs) are being explored to enhance systems' natural language interaction, scene understanding, and human-like reasoning capabilities, but their potential in generating and analyzing complex dynamic scenes remains to be further explored. Visual language models (VLMs) and multimodal large language models (MLLMs) enhance systems' semantic understanding of the environment by fusing visual and language information, providing the possibility for more refined scene parsing. In addition, diffusion models (DMs) and world models (WMs) show great potential in generating high-quality simulation scenes and predicting future states, and are important tools for building high-fidelity virtual test environments and conducting large-scale security verifications[5]. Although these models have driven technological advancements in their respective fields, how to systematically apply them to generate and evaluate driving scenarios covering long tails and extreme situations for comprehensive testing and enhancing system robustness remains a frontier and challenge in current research.

3.2 Input Layer: The Foundation for Enhancing Data and Model robustness

Model robustness begins with high-quality, diverse data input. The traditional principle of minimizing empirical risk can lead to overfitting and insufficient generalization ability of the model. To address this, researchers employed multiple strategies to enhance model stability at the input level.

3.2.1 Adding noise

Generally, we train the model using the Emperical Risk Minimization (ERM) principle:

$$\min_{\theta} L(f)=\frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i),y_i) \quad (1)$$

Let denote the model architecture and its parameters. While minimizing the loss can be achieved by simply memorizing training samples, this leads to poor generalization—the ultimate goal of having the model perform well on unseen data. Models trained via Empirical Risk Minimization (ERM) often fail to perform robustly on distributions slightly different from the training set, a result of overfitting, where the model becomes overly specialized to the training data. To address this, data augmentation is commonly employed. It expands the training set by applying realistic perturbations, such as rotation or scaling of images, forcing the model to learn more essential and invariant features, thereby significantly improving its adaptability to novel samples[6].

3.2.2 Adversarial training

Another key technique is adversarial training. The approach aims to enhance model robustness by actively constructing and defending against "adversarial samples" (samples that add tiny perturbations that are imperceptible to the human eye to clean inputs but can cause the model to make incorrect predictions).

Adversarial training can be written as follows:

$$\min_{\theta} E_{(x,y)\sim D}\left\{\max_{|\delta|_\infty\le\epsilon} \mathrm{L}\left(f_\theta(x+\delta),y\right)\right\} \quad (2)$$

Its core objective is to minimize the model's risk on adversarial samples. Mainstream methods such as fast Gradient (FGM) and Projected Gradient Descent (PGD), as shown in Figure 2, force the model to maintain predictive

consistency[7].

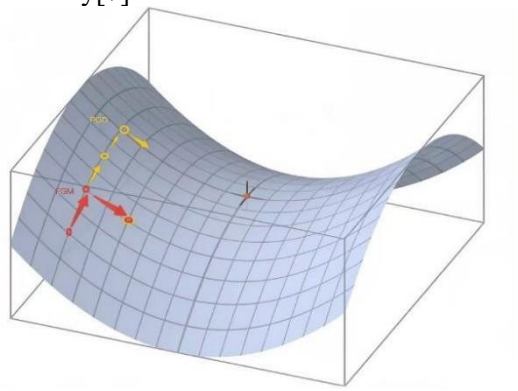

Figure 2. Analysis of FGM and PGD models

This not only enhances the model's defense against malicious attacks, but also serves as an effective regularization tool to improve the model's generalization performance on out-ofdistribution samples.

3.3 Hidden Layers: Regularization and Structural Optimization

At the internal structure level of the model, regularization techniques are crucial for preventing overfitting and enhancing generalization ability. Among them, Dropout is a widely used and highly efficient regularization method.

The calculation formula for the forward propagation of a neural network without DropOut can be described as: the input vector of the l+1 layer is the weight of l+1 multiplied by the output vector of the l layer plus the bias of the l+1 layer. The output vector of the l+1 layer is the input vector of the l+1 layer after the activation function.

$$z_i^{(l+1)}=w_i^{(l+1)}y^l+b_i^{(l+1)} \tag{3}$$

$$y_i^{(l+1)}=f\left(z_i^{(l+1)}\right) \tag{4}$$

The calculation formula for the forward propagation of a neural network with DropOut added can be described as: compared to the previous output vector passing through the Bernoulli distribution, it is similar to passing through a gate filter. " represents the dot product.

$$r_j^{(l)}\sim \text{Bernoulli}(p) \tag{5}$$

$$\widehat{y^{(l)}}=r^{(l)}* y^l \tag{6}$$

$$z_i^{(l+1)}=w_i^{(l+1)}\widehat{y^l}+b_i^{(l+1)} \tag{7}$$

$$y_i^{(l+1)}=f\left(z_i^{(l+1)}\right) \tag{8}$$

The core idea is to randomly "discard" (that is, temporarily shield) some of the neurons and their connections in the hidden layers of the neural network during training[8][9]. Each training iteration is equivalent to being carried out on a randomly generated, "thinner" subnetwork. This process forces the network not to rely on any single neuron or specific combination of local features, thereby encouraging the network to learn more robust and generalized feature representations, effectively alleviating the complex co-adaptation relationships between neurons and enhancing the overall stability of the model.

3.4 Output Layer: Optimizing Learning Objectives and Alleviating Overconfidence

The design of the output layer is directly related to the learning objectives and final performance of the model. In classification tasks, the standard cross-entropy loss function can easily drive the model to become "overconfident" about the prediction results, that is, the probability of predicting the correct category approaches 1 while completely ignoring other categories. This overconfidence reduces the model's calibration and makes it vulnerable when dealing with fuzzy or difficult samples. Two approaches are proposed to address such problems.

3.4.1 Label smoothing

Label Smoothing introduces a certain degree of uncertainty to model training by softening real "hard" labels, such as adjusting 1 in a single-hot encoding to a value slightly less than 1 and evenly distributing the remaining probabilities to other classes.

Let's first look at the formula for Label Smoothing, and before introducing it, let's take a look at the formula for the traditional one-hot encoding as follows:

$$\begin{cases}1, if(i=y)\\0, if(i\neq y)\end{cases} \tag{9}$$

Label Smoothing introduces a factor mechanism, and the formula changes as follows:

$$\begin{cases}(1-factor), if(i=y)\\\frac{factor}{Total\ \ categories}, if(i\neq y)\end{cases} \tag{10}$$

This enables the model not to overfit the absolute labels of the training samples, making the prediction results more "gentle", thereby enhancing the model's generalization ability and robustness[10].

3.4.2 Focal Loss

Another approach is to adjust the loss function itself to address class imbalance or differences in sample difficulty. Focal Loss, by introducing an adjustable modulation factor, dynamically reduces the loss weight of easily classifiable

samples during training and focuses the training on those that are difficult to classify.

Cross-entropy can be written as:

$$l(x,y)=-\log p_k(x) \quad (11)$$

Although it was intended to address the imbalance of positive and negative samples in object detection, its effect can also alleviate the over-optimization of the model for a large number of simple samples, prompt the model to learn more discriminative features, and indirectly enhance the model's robustness on challenging samples.

3.5 Chapter Summary

Machine learning provides powerful perception, decision-making, and generative capabilities for unmanned driving system, but it faces severe robustness challenges in actual deployment. This chapter reviews the role of machine learning in driverless vehicles at the model architecture (LLMs, VLMs, DMs, etc.) and application levels And systematically review a series of key techniques aimed at enhancing model robustness and generalization from the input layer (data augmentation, adversarial training), the hidden layer (Dropout regularization) to the output layer (label smoothing, focus loss). These techniques, through different mechanisms, work together to enable the model to maintain stable and reliable performance in the face of data noise, distribution offsets, adversarial attacks, and rare scenarios, laying the algorithmic foundation for building safe and reliable unmanned driving system. The following chapters will build on this and further explore robustness enhancement schemes for specific segments of driverless vehicles.

4. Comparative Analysis of Robustness Enhancement Techniques and Integration Pathways

4.1 Performance and Challenges of Robustness at All Levels of Driverless Technology

The robustness requirements of unmanned driving system span the entire chain of perception, decisionmaking, and control, with different focuses on performance and challenges.

4.1.1 Perception layer: multi-source fusion and adaptation to extreme environments

The robustness of the perception layer is reflected in the stable perception of uncertain environments and disturbances. The core challenge lies in overcoming the limitations of a single sensor and performance degradation in extreme environments. This is manifested as:

(1) Multi-sensor fusion robustness: The key point is whether the system can maintain overall perception accuracy through complementary and cross-validation of other modal data such as radar and lidar when some sensors (such as cameras overexposed in strong light) fail or their performance declines, and avoid "one loss, all loss".

(2) Adaptability to extreme environments: The system needs to maintain effective perception in harsh conditions such as heavy rain, fog, and night. This requires algorithms not only to handle degraded data such as image blurring and point cloud decay, but also to be able to adaptively adjust the confidence weights[11].

4.1.2 Decision-making level: dynamic adaptation and uncertainty handling

The robustness of the decision-making level requires that the system can still make safe and reasonable driving decisions when the information is incomplete or there are conflicts. The main challenges include:

(1) Dynamic and long-tail scene handling: In the face of rare "long-tail scenes" in the training data (such as animal intrusions, temporary roadblocks), the system needs to avoid decision failure or overconservatism due to the model overfitting common scenarios.

(2) Uncertainty in multi-agent interactions: When interacting with human drivers, pedestrians, etc., accurately predicting the intentions of the other party and making robust game decisions is a huge challenge, and the uncertainty of the prediction model needs to be overcome.

4.1.3 Control layer: perturbation suppression and fault tolerance

The robustness of the control layer is reflected in the vehicle dynamics system's ability to tolerate and stabilize external disturbances and actuator failures.

(1)Dynamic condition adaptation: Under changing conditions such as low adhesion road surfaces

and strong crosswinds, the control system needs to maintain trajectory tracking accuracy and vehicle stability to avoid losing control due to model mismatch.

(2)Actuator Failure response: When critical actuators such as brakes and steering have partial failures, the system must be able to be

restructured through redundant mechanisms or control algorithms to maintain basic controllability and achieve "fail-operable".

4.2 Comparative Analysis of Key Technologies for Enhanced robustness

To address these challenges, the industry has proposed a variety of enhancements at the system architecture, development process, and algorithm levels.

4.2.1 Redundant design and fault-tolerant architecture

This approach implants robustness at the level of system hardware and software design, with the core idea of avoiding single points of failure through backup and diversity.

Hardware redundancy: Multiple configurations of key sensors such as lidar, cameras, computing units (ECUs), and power supplies. When the primary unit fails, the backup unit can take over seamlessly or downgrade. The advantage is that it can directly deal with physical hardware failures and has high reliability; But the cost is a significant increase in cost, power consumption and system complexity.

Software and algorithmic redundancy:

(1) Primary/backup/heterogeneous algorithms: Deploy multiple sets of decision or control algorithms based on different principles, such as rule-based and learning-based. While the primary algorithm is running, the standby algorithm is used for monitoring or parallel computing verification. Switch immediately as soon as the primary algorithm outputs an anomaly or exceeds the safety boundary. This enhances the ability[12] to deal with software flaws and unknown scenarios.

(2) Fault-tolerant control algorithm: When designing the control law, factors such as actuator failure and model uncertainty are taken into account, so that the system performance degrades but remains stable in the event of partial failure. These methods do not add hardware, but rely on precise fault diagnosis and complex algorithm design.

(3) Comparative summary: Redundant design (especially hardware) provides a fundamental guarantee of high reliability and is a common requirement for functional safety (such as ISO 26262). But its "additive" strategy poses cost challenges. Software redundancy and fault-tolerant controls, which compensate for hardware deficiencies with "algorithmic" capabilities, are more flexible but difficult to verify. The future trend is software-hardware collaboration for lightweight redundancy.

4.2.2 Specialized training and simulation testing for robustness

This path is designed to proactively enhance the robustness of the core of the algorithm by exposing the system to a large number of difficult, rare scenarios.

(1) Core Approach: Systematically generate and test scenarios such as extreme weather, sensor failures, and complex traffic conflicts using high-fidelity simulation platforms like CARLA. Identify weak points by quantifying the performance degradation of the algorithm in the simulation (such as reduced detection rate, increased trajectory error).

(2) Key technologies:

Scene generalization and variation: Automatically generate massive, diverse test scenarios through parametric randomization, covering long-tail distributions.

Simulation-reality migration: By means of domain randomization, sensor physical model refinement, etc., narrow the gap between simulation and reality and enhance the credibility of test results.

Parallel accelerated testing: Utilizing parallel computing to significantly enhance testing efficiency in extreme scenarios.

Advantages and limitations: The method is cost-effective, highly secure, and highly repeatable, and is an essential part of verifying and iterating the robustness of algorithms. Its performance is highly dependent on the fidelity of the simulation environment, and good performance in the simulation does not necessarily equate to robustness in the real world. It is often combined with adversarial training, where adversarial perturbations or difficult samples are actively added to force the model to learn more robust feature representations.

4.2.3 Data augmentation and algorithm optimization for robustness

This approach enhances generalization and anti-interference capabilities directly from the model and data levels of machine learning.

(1) Data augmentation: Expand the training data by means of geometric transformations (rotation, clipping), color perturbation, simulated noise (rain, snow, fog), etc., to enhance the model 's tolerance for input variations. This is a fundamental and effective approach, but simple augmentation offers limited improvement to extreme adversarial perturbations.

(2) Adversarial training: Add adversarial samples (data of tiny perturbations that are carefully crafted to cause the model to fail) to the training set to teach the model to resist such attacks. This is one of the most effective ways to enhance a model's adversarial robustness, but it can also result in a slight drop in standard accuracy on clean data and high computational overhead.

(3) Robust Model Architectures and Loss Functions:

Stability regularization: Introducing constraints such as Dropout and label smoothing during training to prevent the model from being overly confident in the training data and improve generalization.

Robust Loss function: Use something like Focal Loss to make the model focus more on hardto-classify samples, or design a loss function that is insensitive to outliers.

(4) Contrastive analysis: This type of method enhances the robustness of the algorithm by nature, does not rely on additional hardware, and is a research hotspot. The challenge is that there may be trade-offs between different augmentation techniques, such as adversarial training and data augmentation, and that augmentation designed for a specific type of interference, such as weather, may be ineffective against other types of interference, such as adversarial attacks. Techniques need to be selected and combined for specific robust threat models.

4.3 Summary of This Chapter

At present, the enhanced robustness of unmanned driving system has formed a multi-level technical system: redundant architectures provide the "hard" guarantee of failure tolerance; Simulation tests and adversarial training form a "drill" mechanism that actively exposes and reinforces weaknesses; Data augmentation and robust algorithms are the "inner strength" that enhances the model's inherent generalization ability.

However, existing methods still face the following core problems:

(1) Trade-off between effect and cost: Hardware redundancy significantly enhances reliability but increases cost and power consumption; Complex robust algorithms, such as adversarial training, bring computational and time costs. How to achieve high-performance robustness on resource-constrained vehicle-mounted platforms is the key to engineering implementation.

(2) Evaluation Criteria and validation gap: Lack of a unified, quantifiable benchmark for robustness evaluation. The fidelity problem of simulation tests makes it difficult for the robustness verified in the simulation to be fully equivalent to real road performance, creating a "verification gap".

(3) Complexity of the combination of techniques: There may be conflicts or coupling effects among different robustness techniques (such as adversarial training and specific data augmentation), and how to systematically and collaboratively optimize multiple techniques rather than simply stack them remains an open question.

(4) Generalization for unknown threats: Existing methods are mostly aimed at known, modelable disturbances (such as specific weather, preset attacks). How systems can maintain "resilience" against unknown, out-of-distribution new threats is a higher-dimensional challenge.

Future research will focus more on lightweight, explainable, and certified robustness approaches, driving the evolution from dealing with known threats to resilient systems with greater adaptability and resilience, and accelerating the development of reliable and trustworthy driverless technology by building more open benchmarking platforms.

5. Summary and Outlook

5.1 Summary

This paper provides a systematic review and study of the robustness issues of machine learningbased unmanned driving system. By analyzing the technical architecture of unmanned driving system and the core challenges they face in practical applications, we have identified robustness-the ability of a system to operate safely, stably, and reliably in the face of uncertainty, disruption, and extreme conditions-as the key bottleneck restricting the technology from research to large-scale commercial deployment.

At the problem definition level, this paper first sorts out the core technology stack of "perceptiondecision-making-control" for unmanned driving system. The study points out that the robustness challenge runs through the entire technical process: at the perception level, it is mainly manifested as vulnerability to adversarial attacks and sensitivity to environmental disturbances; At the

decisionmaking level, the core challenge lies in dealing with the failure of long-tail scenarios and the uncertainty of multi-agent interactions; At the control level, there are limitations in actuator dynamic response and the risk of fault propagation. In addition, from a data and system-wide perspective, issues such as the simulation-reality gap, coupling faults between modules, and cybersecurity threats further constitute the full-chain robustness challenge spanning "data-algorithm-hardware-system".

At the technical response level, this paper constructs a methodological system for enhanced robustness from two dimensions:

(1) Internal enhancement at the machine learning algorithm level: Focused on the core approach starting from the model training process. At the input layer, through data augmentation and adversarial training, actively construct and learn to respond to perturbations, enhancing the model's adaptability to noise and out-of-distribution samples. In the hidden layer, regularization techniques such as Dropout are used to prevent overfitting and encourage the learning of more generalized features. In the output layer, apply label smoothing and focus loss to alleviate the "overconfidence" of model predictions and optimize the processing of difficult samples. These techniques work together to enhance the model's own inherent robustness.

(2) External safeguards at the system architecture and validation level: Systematically sorting out the crucial enhancement paths in engineering practice. Redundant design (including hardware, software and algorithmic redundancy) provides a "safety bottom line" for failure tolerance architecturally and is the foundation for meeting functional safety standards. Robustness tests based on high-fidelity simulation (represented by platforms such as CARLA) build an efficient and secure virtual verification environment that enables active exposure and quantitative analysis of algorithmic weaknesses through systematic generation and evaluation of extreme and rare scenarios, serving as a key bridge between algorithmic innovation and engineering implementation.

To sum up, improving the robustness of unmanned driving system is a complex systems engineering project that requires a close combination and coordinated optimization of the inherent robustness of the algorithm, the fault tolerance of the system architecture, and the completeness of the verification throughout the entire process. The current research and practice have made significant progress along these paths, but there is still a significant gap from building fully unmanned driving systems that are highly reliable in any open environment.

5.2 Future Prospects

Based on the research in this paper, we believe that future research on the robustness of unmanned driving system will face opportunities and challenges in the following directions:

(1) Evolution from "robustness" to "system resilience" : Future research should move beyond static defense against known, specific threats and towards building "resilient systems" with online adaptation and autonomous recovery capabilities. This requires exploring technologies such as meta-learning and online continuous learning to enable systems to autonomously detect, diagnose unknown or outof-distribution scenarios, and quickly adjust strategies at runtime. At the same time, developing interpretable uncertainty quantification methods that enable systems not only to output decisions but also to assess their confidence levels is crucial for building human-machine trust and clarifying accountability.

(2) Integration of lightweight and certified robustness technologies: Currently, many advanced robustness methods (such as complex adversarial training and large-scale model integration) have problems of high computational overhead and difficulty in real-time deployment on automotivegrade chips. Therefore, developing computationally efficient lightweight robust algorithms and software-hardware co-design solutions will be key to engineering implementation. In addition, promote the formal verification and security certification process of robust technologies, and establish a full-chain trusted evidence chain from theoretical assurance, simulation testing to real vehicle verification to meet increasingly strict security standards such as ISO 21448 (SOTIF).

(3) Closing the "verification gap" between simulation and reality: Although simulation testing is indispensable, the "verification gap" caused by its limited fidelity remains a core challenge. In the future, more sophisticated sensor physical models, traffic participant behavior models, and vehicle dynamics models

need to be integrated to develop digital twin-level high-fidelity simulation environments. At the same time, there is an urgent need to establish an open and unified multi-dimensional robustness benchmark set (covering adversarial attacks, extreme weather, long-tail interactions, etc.) to provide objective and comparable performance evaluation criteria to drive the healthy development of the domain.

(4) Coordination and system-level optimization of multiple technical paths: A single technology is difficult to solve all robustness problems. Further research will be needed on the interaction, conflict, and synergy mechanisms among different robustness enhancement techniques in the future. For example, how can the robust model features obtained from adversarial training be seamlessly integrated with the failover logic of redundant control systems? How to use the data generated from simulation tests to direct data augmentation strategies or model architecture searches? Achieving robust collaborative design across layers and modules is key to unlocking the overall potential of the system.

(5) Addressing New threats and regulating ethical boundaries: As technology advances, new challenges keep emerging. Swarm intelligence robustness based on V2X collaboration and defense against new types of cyber attacks against multi-sensor fusion layers will be important topics. Furthermore, robustness studies must be closely integrated with ethical norms to ensure that decisions made by systems in inevitable extreme situations (accident critical states) are in line with ethical consensus in human society and provide technical basis for the formulation of relevant laws and regulations.

The road to robustness for unmanned driving system is long and arduous. It requires a deep integration and continuous innovation of multiple disciplines such as computer science, control theory, automotive engineering, cyber security, and ethics. By focusing on these core areas and persevering in tackling scientific and engineering challenges, we are expected to gradually build a solid safety barrier for unmanned driving system, ultimately promoting the safe, reliable and responsible integration of this transformative technology into social life and realizing its grand vision of improving traffic efficiency and ensuring travel safety.

References

[1] Zhang Xinyu; Gao Hongbo; Zhao Jianhui; Zhou Mo; A Review of deep learning-based autonomous driving technology [J]. Journal of Tsinghua University (Natural Science Edition),2018,v.58(04).

[2] An Introduction to Autonomous driving technology [M]. Beijing: Tsinghua University Press, 2019.

[3] Driving forefront. CN. (2025) autopilot robustness is a frequently mentioned what https://zhuanlan.zhihu.com/p/16009629112

[4] Zhou Zhihua. Machine learning [M]. Tsinghua University Press,2016.

[5] The Heart of Autonomous Driving.cn.(2025) TUM latest! Comprehensive combing autopilot based model: LLM/VLM/MLLM/diffusion model and the model for the net. https://mp.weixin.qq.com/s/QlYYiHGpOq2tCSeur01Nfw.

[6] Su Jianlin. Lipschitz constraint in the deep learning: the generalization and the generated model [EB/OL]. (2018-10-07) [2024-01-01]. HTTP: / / https://kexue.fm/archives/6051

[7] Su Jianlin. (Mar. 01, 2020). The confrontation training, discussion, significance, methods and thinking (with Keras implementation) "[Blog post]. Retrieved from https://kexue.fm/archives/7234

[8] Ghiasi G, Lin T Y, Le Q V. Dropblock: A regularization method for convolutional networks[J]. arXiv preprint arXiv:1810.12890, 2018.

[9] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[10] ZhuoBuFan. CN. (2025) Summary of Label Smoothing Techniques. https://zhuanlan.zhihu.com/p/1931870988791976545

[11] Zhang, Y., Carballo, A., Yang, H., Takeda, K. (2023). Article Title. ISPRS Journal of Photogrammetry and Remote Sensing, 196, 146-177.

[12] GoTSHgo. CN. (2025). Dual-machine algorithm: Active-standby, Active-active Comparison. https://blog.csdn.net/goTsHgo/article/details/146341926

