Trustworthy Autonomous Driving with AI: Status, Challenges and Opportunities

Chengcheng Wang, Yuansheng Song, Jiahao Zhang, Liang Zhao, Chen Chen, Xiaoheng Deng, Fangming Liu and Shaohua Wan

Abstract—Leveraging artificial intelligence (AI) technologies, autonomous driving (AD) enables the perception and recognition of intricate surroundings while making intelligent decisions for control. However, limited by complex environments, heterogeneous resources, and real-time requirements, AI-driven AD in complex scenarios often faces key scientific issues such as performance ensuring and decision interpreting, posing significant threats to both public trust in AD and the macroscopic efficiency of the intelligent transportation system (ITS). In this research, we provide a systematic review of recent progress and important contributions to the trustworthiness of AD. First, we introduce the background of AD with AI, including the basic paradigm of AD, the benefits of deep learning (DL) and AI-generated content (AIGC) for AD development, and existing barriers. Afterward, we investigate the trustworthiness of AD from three perspectives: interpretability, robustness, and exactness. Specifically, i) we categorize landmark methods for AD interpretability, including feature attribution, visual explanation, latent representation, and efficient understanding and reasoning of large language models (LLMs); ii) we briefly discuss the robustness of AD under various challenging conditions, such as incomplete and differentiated training data and external attack interference; iii) we focus on representative solutions on AD exactness along architecture and training brought by the differentiated balance between resource provision and performance. Finally, we present a multi-scale trustworthy evolution framework bridging micro-scale ego-vehicle intelligence with macro-scale ITS, systematically mapping relevant challenges to strategic opportunities. This research is expected to provide a fundamental direction for researchers new to the trustworthiness of AD with AI.

Index Terms—Autonomous driving, Artificial intelligence, Interpretability, Robustness, Exactness

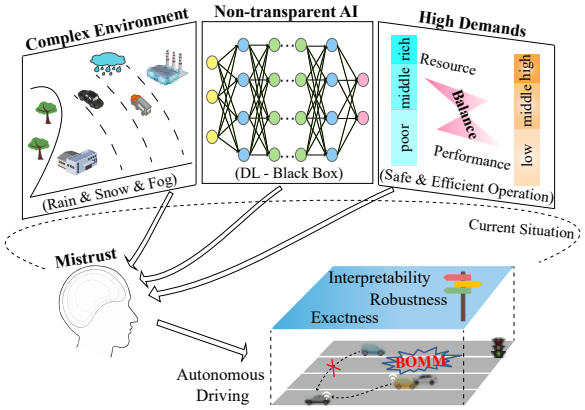


Fig. 1: Efficient and trustworthy AD requires comprehensive consideration of factors such as environmental complexities, algorithmic non-transparency, and computational trade-offs between onboard resources and sensitive performance demands. Peeking into the robustness of AD models under various harsh conditions, improving the exactness of AD in decision-making tasks, and enhancing the interpretability of AD in operations can provide effective performance assurance and failure analysis for AD systems. Such features play a crucial role in strengthening public trust in AD.

I. Introduction

AUTONOMOUS driving (AD), as a result of the integration of cutting-edge technology with the automotive manufacturing and transportation industries, holds significant promise for improving road safety, providing novel driving experiences, and enhancing the efficiency of social operations [1]. According to the latest statistics, the global market value of autonomous vehicles (AVs) is expected to reach $5,439.46 billion by 2035, an approximately 19.9-fold increase from $273.75 billion in 2025 [2]. This indicates that worldwide AD is progressing rapidly. However, as a safety-critical real-world application, the trustworthiness of AD with artificial intelligence (AI) has recently been a matter of public concern, especially in light of the high frequency of traffic accidents [3]. So, what causes AD incidents? What malfunction in the AD system causes the impact or collision? This line of reasoning naturally raises the following question: How can AD be made trustworthy by humans?

As illustrated in Fig. 1, the complex dynamic road environment, demanding real-time computational requirements, and algorithmic opacity highlight the challenges to the trustworthiness of AD. Crucially, micro-level trustwor-

Chengcheng Wang, Yuansheng Song, Jiahao Zhang, and Shaohua Wan are with the University of Electronic Science and Technology of China, Chengdu, China. They are also with the Shenzhen Institute for Advanced Study, UESTC, Shenzhen 518110, China (e-mail: wccwb1@outlook.com; 1782751638@qq.com; zjh15539005762@gmail.com;shaohua.wan@uestc.edu.cn).

Liang Zhao is with the School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China (e-mail:lzhao@sau.edu.cn).

Chen Chen is with the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: cc2000@mail.xidian.edu.cn).

Xiaoheng Deng is with the School of Computer Science and Engineering, Central South University, Changsha 410075, China (e-mail: dxh@csu.edu.cn).

Fangming Liu is with Peng Cheng Laboratory, and Huazhong University of Science and Technology, China (e-mail: fangminghk@gmail.com).

Shaohua Wan is the corresponding author.

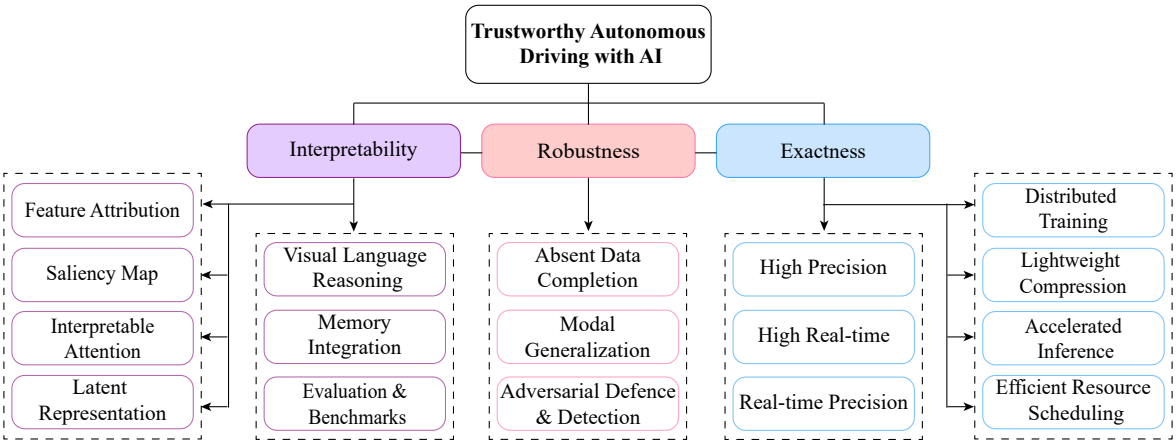


Fig. 2: Structure of the overall research.

thiness dictates macroscopic transportation performance: ego-AV failures amplify traffic shockwaves, whereas reliable AI-driven AVs smooth mixed-traffic flow and enhance network throughput [4]–[7]. Given this profound system-wide impact, establishing the trustworthiness of AI-driven AD has become an imperative priority for both academia and industry [8]. Therefore, to cultivate highly trustworthy AD that truly empowers ITS, all dimensional indicators of the AD system should be fully ensured and validated. Among these trustworthiness measures, interpretability, robustness, and exactness stand out as key evaluation dimensions, each with its significance:

Interpretability, an important adjunct for trustworthiness, is vital for AD because modern architectures heavily rely on opaque deep neural networks (DNNs) and hallucination-prone large language models (LLMs) [9]. Operating end-to-end, such systems introduce severe vulnerabilities in safety-critical edge cases, where the inability to audit the causal logic of behavioral anomalies leaves AD legally unaccountable [10]. Furthermore, global regulations such as the General Data Protection Regulation (GDPR) have also explicitly demanded algorithmic fairness and transparency alongside efficiency [11]. As a result, to satisfy both regulatory mandates and road users' expectations, enhancing interpretability is an urgent imperative to address these challenges and ultimately bolster public trust in AD. At the system level, interpretable AD fosters predictable interactions in mixed-traffic environments, thereby reducing traffic conflicts and improving the overall throughput of the road network.

Robustness is another key factor for fostering public trust in AVs, as it guarantees reliable operation across increasingly diverse scenarios [12]. However, the reliability of AD systems remains persistently challenged by the intricate interplay of complex traffic dynamics, architectural limitations, and adversarial interference. In particular, inherent 'blind spots' resulting from incomplete data modalities, long-tail distributions, and domain shifts introduce unpredictable failure modes that remain resistant to traditional optimization techniques [13]. Consequently, the capacity of AVs to sustain optimal performance under limited perception and within harsh, real-world edge cases (e.g., extreme weather or anomalous road conditions) is not merely a technical metric; it is a critical prerequisite for aligning AD systems with public trust. Macroscopically, robust AD avoids unpredictable disengagements and sudden braking, which are primary triggers for phantom traffic jams and network-wide congestion.

Exactness serves as a foundational cornerstone for AD trustworthiness, demanding high-fidelity execution with rigorous accuracy and minimal latency across perception, prediction, and decision-making [14]. Nevertheless, the relentless expansion of model complexity often conflicts with onboard resource constraints, forcing a critical trade-off between computational overhead and precision [15]. To address this challenge, it is imperative to rigorously characterize and maintain AD exactness, as any divergence from precise control directly undermines the safety guarantees upon which public confidence relies.Furthermore, from a system perspective, such precise execution is indispensable for advanced cooperative driving applications, such as vehicle platooning and vehicle-to-everything (V2X) enabled intersection scheduling, which can significantly multiply road capacity.

While existing research has used robustness and interpretability as key pillars of trustworthiness [16], no published study has yet comprehensively generalized these aforementioned issues to AD with AI. To bridge this gap, we present this research and further highlight the following contributions.

- This research reviews the key benefits and existing challenges that AI technologies represented by DL and AIGC have played in the historical development of AD, and highlights the importance and necessity of research into trustworthy AD with AI.
- To the best of our knowledge, this is the first research to verify the trustworthiness of AD from the three perspectives of interpretability, robustness, and exactness of AD to enhance public trust, with an in-depth exploration of the internal working mechanism of AD with AI.
- We provide a systematic analysis and comparison

of various AI categories that open new avenues for AD systems. Complementing this, we propose a multi-scale trustworthiness framework for AI-based AD, extending from ego-AV intelligence to ITS, and outline key challenges and insights regarding its trustworthiness.

The remainder of this research is organized as follows. Section II provides a brief overview of AD with AI. Section III-V review the interpretability, robustness, and exactness of AD. Section VI identifies open challenges and potential directions. Finally, Section VII concludes the research. The organization of this research is illustrated in Fig. 2.

II. Background

This section provides a brief overview of two paradigms in the history of AD development, as well as the driving forces and potential challenges that AI technologies, represented by traditional DL and emerging AIGC, bring to AD.

A. AD

Looking back at AD's evolution, two paradigms have dominated: modular and end-to-end [17]. As shown in Fig. 3, traditional AD systems employ a modular architecture that excels in interpretability and localized debugging. However, this unidirectional, cascaded pipeline suffers from degraded accuracy and efficiency in complex scenarios due to three architectural bottlenecks: (i) Cascading Error Propagation, where minor upstream errors compound sequentially downstream without closed-loop feedback; (ii) Information Representation Bottlenecks, where compressing high-dimensional environmental data into low-dimensional intermediate states filters out critical context for edge cases; and (iii) Optimization Objective Misalignment, wherein modules are trained in isolation via surrogate loss functions rather than global driving metrics.

In contrast, end-to-end AD uses raw sensor data as input and incorporates the vehicle's functional processes such as perception, planning, and decision-making into a co-trainable model, aiming to optimize the final task, mitigate error accumulation, and improve computational efficiency. However, it is currently limited to handling simple driving tasks and severely lacks interpretability. Against this background, it is crucial to provide further theoretical proof for the safety and trustworthiness of AD, such as a deeper understanding of scenarios, efficient decision-making, and an explanation of specific network behaviors.

B. AD with DL

As a key enabler for AD, AI techniques, especially DL, have always been essential to AD [18]. For instance, convolutional neural network has replaced early manual feature representations and shown impressive performance in visual perception tasks [19]. While traditional DL solutions have achieved remarkable results in AD development, most can only rely on existing multi-modal data for subsequent operations [20]. There are still substantial technical barriers to large-scale high-quality data requirements, universal model generalization, and reliable performance guarantees.

Specifically, (i) applying DL to AD requires large-scale, high-quality training data, which is costly to collect, and the training typically consumes a lot of energy and computing resources; (ii) DL models are prone to overfitting, leading to limited generalization, particularly regarding the long-tail issue of AD; (iii) catastrophic forgetting and susceptibility to adversarial attacks further challenge the efficiency and robustness of AD; (iv) moreover, the black-box nature of highly non-linear DNNs hinders human understanding of decision-making, raising major ethical and legal concerns [10]. Therefore, exploring trustworthy AI solutions or enhancing DL's robustness and comprehensibility is essential to strengthen AD systems.

C. AD with AIGC

AIGC as a breakthrough AI paradigm generates innovative content by learning from massive datasets, undergoing a human-like process evolving from imitation to creation [21]. Research indicates that integrating AIGC into AD systems significantly enhances driving experience and safety. Specifically, by synthesizing data from on-board sensors, road conditions, and maps, AIGC achieves a comprehensive understanding and accurate prediction of AD scenarios [22]. Furthermore, through collaborative optimization of control and user interaction modules, AIGC facilitates efficient execution of deriving behaviors, thereby enhancing the safety, stability, and intelligence of AD.

Despite AIGC's potential for AD, trustworthiness remains a critical obstacle [23]. First, AIGC's hallucinatory nature leads to low model transparency, making the rationale behind its generated content and decisions difficult to explain, which weakens AD decision trustworthiness [24]. Second, AIGC models are vulnerable to adversarial attacks during training and fine-tuning; unclean data severely damages downstream AD tasks relying on these foundation/generative models, compromising AD stability and robustness. Lastly, high resource consumption throughout the AIGC lifecycle challenges resource-constrained edge devices in ensuring precise, efficient, and time-sensitive AD execution. Together, these factors significantly impact public trust in AD.

III. Interpretability for AD Trustworthiness

To enhance public trust, AD systems need to provide a trustworthy explanation for their actions. This section outlines recent advances in traditional interpretable methods and innovative LLMs for AD interpretability. The classification of interpretable AD methods in this research is presented in Fig. 4 and representative studies are illustrated in Table I.

TABLE I
Representative research that can be used to interpret the internal working mechanisms of AD (Sec. 3).

Ref. No.	Interpretable technique(s) for AD				Interpretable Operation(s) for AD			
	feature attribution	visual explanation	latent representation	LLM	perception	decision	planning	control
Interpretable-ICCV'17 [25]		✓						✓
Visualbackprop-ICRA'18 [26]	✓							✓
Grid saliency-NeurIPS'19 [27]		✓			✓			
Explaining-CVPR'20 [28]		✓					✓	
Explainable-CVPR'20 [29]				✓	✓	✓		
Drive-ICCV'21 [30]		✓				✓		
NEAT-ICCV'21 [31]			✓			✓	✓	
Learning-ICRA'21 [32]			✓				✓	
Multitask-CVPR'21 [33]			✓					✓
Interpretable-TITS'21 [34]			✓		✓			
ThirdEye-ASE'22 [35]		✓				✓		
PlanT-CoRL'22 [36]		✓					✓	
GPTDriver-NeurIPS'23 [37]				✓			✓	
Driving-ICRA'24 [38]				✓	✓	✓		
Deepsecdrive-IS'24 [39]	✓					✓		
Driving-WACV'24 [40]		✓				✓		
VLAAD-WACV'24 [41]				✓	✓			
RAGDriver-RSS'24 [42]				✓	✓	✓	✓	✓
DiLu-ICLR'24 [43]				✓		✓		

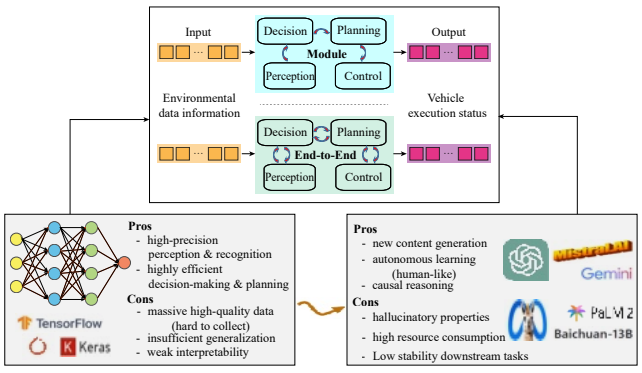


Fig. 3: The developmental paradigm of AD and the impact of traditional DL and emerging AIGC during its evolution.

A. Traditional Interpretable Methods for AD

Self-enhancing AD models and post-AD explanations are two primary approaches to improve AD interpretability [44]. The former continuously improves explanation capabilities through learning and optimization to elucidate internal mechanisms, whereas the latter leverages post-hoc insights like saliency maps from trained models. Below, we briefly review three major categories of interpretable AD methods.

Feature Attribution. As a model-agnostic technique, feature attribution quantifies the contribution of input features to AD outputs, offering finer-grained interpretation. Notably, the shapley additive explanation (SHAP) stands out as the only expectation-based feature attribution method that is both locally accurate and consistent [45]. It has been widely used to assess feature importance in AD; for instance, [39] used SHAP to quantify feature impacts for local and global decision explanations, while [46] employed it to analyze critical attributes influencing road accidents for prevention insights. Another classic approach is gradient-based attribution [26], [47], [48]. For example, [49] introduced RestoreX-AI under attention maps to evaluate restored images by weighting loss, improving object detection in adverse weather. Generally, gradient-based methods reflect input importance by analyzing the gradient of network outputs relative to inputs, thereby explaining AD behaviors.

While valuable, feature attribution methods like SHAP and gradients become highly unstable when handling nonlinear, high-dimensional data. They are prone to gradient explosion or vanishing issues and lack a comprehensive analysis of complex, cross-dimensional feature interactions. This presents a significant challenge for AD systems requiring large-scale data processing and high-temporal assurance.

Visual Explanation. AD advancement heavily relies on DNNs for processing visual streams,making visual interpretability crucial. Two primary tools, attention and saliency maps, are often combined to enhance understanding.

a) Saliency map is the most widely used post-AD explanatory technology, highlighting critical image regions to visualize how input data impacts AD behavior [27]. For instance, [50] mapped AD traffic entity behaviors into a Markov decision process to forecast waypoints, utilizing saliency maps to explain internal reasoning and avoid AD collisions. Mohseni et al. [51] used VisualBackProp [26] to generate teacher model saliency maps, training the student model to predict the failure probability of AD's steering control. Similarly, [52] utilized saliency maps to guide generative adversarial networks, producing informative counterfactual explanations near AD decision boundaries. However, saliency maps only feature input-level highlights without explaining the underlying decision-making mechanism. Consequently, they offer relatively weak interpretability, and their effectiveness remains difficult to

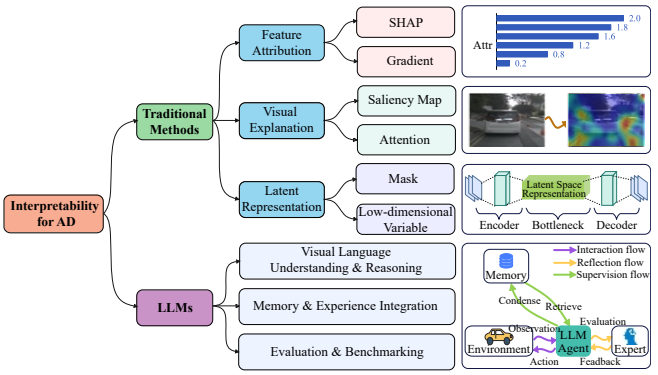


Fig. 4: Summary of methods for interpretable AD. They enable humans to gain a deep and thorough understanding of the internal mechanisms of AD behavior, thereby enhancing trust in AD systems.

assess [53].

b) Interpretable attention stems from learned attentional weights learned within neural networks, highlighting specific features or regions to provide visual interpretations of local effects on AD decisions [25]. Several studies have utilized attention weights to adaptively aggregate important features from intermediate feature maps, pooled from either fixed grids or diverse regions of interest [28], [54]. For example, DRIVE [30] simulates top-down and bottom-up visual attention to highlight critical areas for driving decisions. Thirdeye [35] converts attention maps into confidence scores to predict and distinguish safe versus unsafe driving behaviors. Recent Transformer advances have further enhanced AD interpretability. Specifically, [40] converted image features into concept scores and employed Transformer attention to determine when to display automated explanations to the driver. PlanT [36] proposed an object-level Transformer along with an evaluation protocol to interpret AD decisions. Indeed, while the attention mechanism provides valuable insights, it is typically model-specific and can capture nuances diverging from human intuition, thereby affecting AD trustworthiness. More importantly, [55] challenges this paradigm by demonstrating that learned attention mechanisms do not necessarily increase neural network transparency.

Interpretable Latent Representation. Complex driving environments often render AD inputs high-dimensional and opaque. Moreover, because human driving behavior is driven by unobservable latent factors, mapping complex data into low-dimensional, meaningful latent spaces has become a crucial strategy to enhance AD decision interpretability [32]. For example, [56] compressed perceptual information and internal states into a latent space to extract low-dimensional semantic representations for decision interpretation. Similarly, [34] introduced a sequential latent environment model whose generated semantic bird's-eye-view masks provides automatic explanations of environment understanding; however, its model-free policy learning lacks insight into the model underlying decision-making mechanics. Additionally, [57] applied evidential DL to map epistemic uncertainty to low-dimensional latent variables for trajectory prediction. Although this approach requires only a single forward pass without out-of-distribution training, its slow training process limits the latent space's representational capacity. To enhance safety, [58] introduced an interpretable sensor fusion Transformer that utilizes target density maps and traffic rules from the encoder to constrain actions within safe sets.

While these methods offer valuable insights, they remain susceptible to the interpreter's subjective bias. Furthermore, most existing works are limited to specific auxiliary AD tasks, rather than providing a comprehensive, end-to-end quantitative analysis.

B. LLMs for interpretable AD

As a natural vehicle for human comprehension, natural language is a crucial catalyst for building trust in future AD. Unlike traditional DNN-based interpretable methods, LLMs excel in natural language processing, commonsense reasoning, and human-machine interaction, bringing revolutionary possibilities for AD transparency [66]. Specifically, by fusing their inherent commonsense with multi-source data, such as onboard sensors, road conditions, and maps, LLMs can deeply comprehend AD scenarios and narrow the knowledge gap between AD systems and human experts, thereby enhancing driving safety and trustworthiness [67]. Consequently, LLM-based interpretable AD has emerged as a vibrant and crucial research frontier.

Visual Language Understanding and Reasoning. The exceptional language generation, visual understanding, and reasoning capabilities of LLMs provide essential tools for AD interpretability [38]. For instance, DriveGPT4 [68] and VLAAD [41] introduced visual command adaptation to AD, effectively interpreting AV actions across diverse driving scenarios through fine-tuning. Recent research has extensively explored integrating LLMs directly into AD systems as core modules for decision-making and reasoning to enhance transparency [69]. For instance, GPT-Driver [37] reformulated motion planning as a language modeling task, leveraging LLMs to generate highly interpretable trajectory descriptions by visualizing inference outputs, highlighting key objects, and assessing their potential impacts. [70] integrated LLMs into AD systems to dynamically adjust driving behaviors based on vocal commands, fostering user trust and system interpretability. Collectively, these studies demonstrate that LLMs can serve as high-level decision engines to significantly elevate AD interpretability.

Memory and Experience Integration. A key advantage of human driving is the continuous accumulation of experience, a capability researchers seek to empower LLMs with [42]. For example, [71] proposed an AD system emulating human driving via environment, agent, memory, and expert modules, demonstrating for the first time LLM's ability to interpret and interact with AD environments.

TABLE II
Driving datasets or benchmarks with additional annotations for evaluating the interpretability of AD systems and the characteristics of each.

Benchmark	Original Dataset	Size	Key highlights
BDD-OIA [29]	BDD100K	23K videos (5s)	∗ object-based interpretation with 4 actions and 21 annotation extensions
DQA [38]	generated by GPT	160K QA pairs	• the same expert LLM grader (GPT-3.5) for performance evaluation
VLAAD [41]	HDD BDD, BDD-X	64K video instruction-following	∗ integrateing visual instruction-tuning into AD ∗ public release of VLAAD weights
BDD-X [54]	BDD	7K videos (40s)	• textual driving description and explanation with timestamps for AVs
Reason2Drive [59]	nuScenes Waymo, ONCE	over 600K video-text pairs	∗ a distinctive chain-based reasoning (perception-prediction-reasoning) ∗ a novel evaluation metric for addressing the semantic ambiguity
HDD [60]	-	104-hour driving	• a novel annotation that decomposes driver behaviors into a 4-layer representation: Goal-oriented, Stimulus-driven, Cause, and Attention
CTA [61]	-	1935 videos (17.7s)	∗ an accident analysis benchmark with time intervals and semantic labels
Talk2Car [62]	nuScenes	850 videos (11959 commands)	• the first object referral dataset with natural language AV commands
DoTA [63]	-	4,677 videos (73,193s)	∗ the first publicly-available When-Where-What pipeline dataset with temporal, spatial, and categorical annotations
Lyft Level5 [64]	-	162K scenes (25s)	• the largest motion prediction dataset available at launch for planning
BDD-A Extended [65]	BDD-A	1103 videos (10s)	∗ a dataset solution for explanation timing, necessity, and content

DiLu [43] incorporates human knowledge into AD through a framework equipping the driving agent with reasoning, reflection, and memory modules. Its memory module stores past driving cases and integrates them with existing knowledge for efficient interpretation. [72] introduced tool libraries and cognitive memories to gather and interpret environmental information; the former extracts data from neural module outputs, while the latter stores driving rules and experiences. Together, these enhancements make AD information collection and decision-making more transparent and explainable. Ultimately, these methods not only boost LLM generalization but also simulate the human driving learning process, providing valuable insights for interpretable AD development.

Evaluation and Benchmarks. To interpret AD behaviors, early efforts focused on annotating datasets for explanation generation [29]. Supervising neural networks with these explanations has proven effective, though acquiring such annotations is challenging and relies on strong assumptions [54]. In the LLM era, researchers have constructed advanced evaluation datasets and benchmarks to enhance AD interpretability. For instance, Talk2Car [62], the first object referral dataset, provides natural language instructions for AVs. Building on this, interpretable reasoning datasets such as DQA [38] and Reason2Drive [59] were proposed to comprehensively evaluate the explanatory and chain-of-thought reasoning capabilities of LLMs in AD scenarios. These benchmarks, which quantify and compare different methodologies while guiding future interpretability research, are summarized in Table II.

In general, LLM unlock new possibilities for elevating AD intelligence and interpretability. However, AD systems cannot fully rely on LLM outputs for final decisions, as LLM hallucinations can easily generate inaccurate information [24]. Furthermore, most existing works remain confined to specific tasks or scenarios, leaving a substantial gap before practical product application. To ultimately achieve a trustworthy and transparent AD system, future research must construct a more durable, universal, and secure LLM-integrated AD architecture while continuously enhancing its generalization capabilities and safety.

IV. Robustness for AD Trustworthiness

To enhance reliability and public trust, AD systems must adapt to harsh and challenging environments. This section reviews research on AD robustness, focusing on the challenges of incomplete and inconsistent driving data, as well as susceptibility to cyber-attacks. Representative studies are presented in Table III.

A. Absent Data Completion

AD systems rely heavily on high-quality environmental data to ensure efficient module collaboration and safe operation. However, real-world challenges, such as sensor limitations, adverse weather, and occlusions, often result in defective or incomplete data, severely compromising road safety [73]–[75]. To address the weak robustness of AD models caused by missing data, this subsection explores the perspective of data completion, as illustrated in Fig. 5. The specific content is elaborated below.

Uni-modal Data Completion. It aims to fill in absent data using single-type data algorithms to achieve more efficient and robust AD. Current methods are primarily statistical-based and DL-based. For instance, [76] proposed a graphics processing unit (GPU)-based IDW interpolation algorithm to quickly interpolate missing road LiDAR data, improving AD robustness against AV occlusion. To handle camera failures,, a plug-and-play MVR was introduced for robust absent view recovery [77]. From a missing data detection perspective, [78] proposed a two-step approach, combining supervised learning for class boundaries with semi-supervised refinement of normal distributions, to identify LiDAR data missing due to rainfall degradation, defining the anomaly score as the distance from the query sample to the normal class.

While uni-modal data completion shows promise in AD robustness, statistical approaches typically require strict priors, significantly limiting model generalization [79]. Meanwhile, DL-based methods often necessitate extensive

TABLE III
Representative studies focusing on enhancing the robustness of AD systems.

Technique(s) \ Scope(s)		Sample	Model
Absence completion	uni-modal	[76], [78]	[77]
	multi-modal	[81]	[82]–[84]
Model Generalization	DA	[85], [86]	[85]–[87]
	ZSL/FSL	[88]	[89]–[91]
	OWL	-	[92]–[94]
Adversarial Attacks	detection	[95]–[97]	[98]
	defense	[99]	[100], [101]

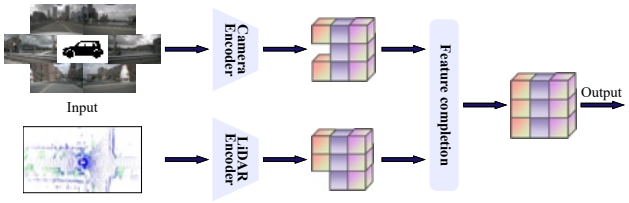


Fig. 5: Interpolating data with complementary LiDAR and camera inputs for robust AD.

training data. These limitations highlight the constraints of single-type sensors, such as the short-range semantic richness of cameras versus the long-range precision of LiDAR, which suffer from data singleness [80]. Therefore, further research into effective imputation for individual data types remains vital.

Multi-modal Data Completion. It interpolates data by exploiting complementary modalities to enhance the reliability and security of AD [82]. It outperforms uni-modal completion by utilizing clear streams for supplementary information. For instance, MetaBEV [83] addresses robustness issues in 3D detection and map segmentation caused by missing sensor signals. UniM^{2}AE [102] facilitates multi-modal masked input reconstruction via a unified 3D volume space. AutoFed [81] uses an inter-modal autoencoder with residual connections to interpolate missing modalities, outperforming zero-padding in precision and recall. To improve flexibility, ST-MVDNet [103] introduces a mutual learning pipeline that prevents over-reliance on a single sensor when LiDAR or radar is partially absent. Similarly, [104] presents a symmetric semantic segmentation architecture that ensures reliable operation if either LiDAR or RGB data is corrupted. For severe data loss, SMIL [84] introduces Bayesian meta-learning, outperforming baselines like autoencoders and GANs.

While many studies improve AD robustness through multi-modal approaches, most rely on complete modalities, with few addressing missing data completion [105]. Additionally, heterogeneous modalities (e.g., images, text, audio) poses challenges in data alignment, synchronization, noise processing, real-time computational efficiency, and data privacy. Further research in these areas is crucial to enhancing AD system robustness.

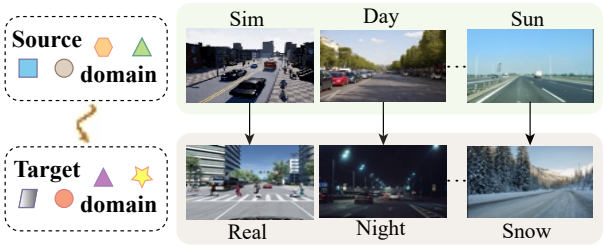


Fig. 6: Applying DA across various scenarios, such as simulation-to-real, day-to-night, and sun-to-snow transitions, enhances AD robustness.

B. Model generalization

Model generalization is crucial for practical AI-based AD, enabling AVs to perform effectively across diverse and dynamic real-world domains [106]. However, real-world data often exhibits significant environmental discrepancies, and data annotation incurs substantial costs [107]. To address these challenges, this subsection reviews key techniques aimed at enhancing the robust generalization of AD models.

Domain Adaptation (DA). It aims to transfer behavioral capabilities learned by AD models from a known source domain to a different but related target domain by narrowing the domain gap (i.e., inconsistent data distributions) [108], as illustrated in Fig. 6. Recent studies have focused on environment transitions (e.g., sun-to-rain/snow/fog) to enhance AD robustness [109], particularly addressing extreme weather conditions by employing specific strategies like GAN-based synthetic weather augmentation, image denoising, and comprehensive all-weather data training [110], as well as knowledge amalgamation [111] for reliable object detection. Below is a common classification of DA methods.

a) Self-supervised DA (Self-DA) enables AD models to learn task features even in unlabeled datasets [87]. For instance, [112] improves visual navigation in unseen environments via global map consistency, whiSSTN [113], a DNN-based self-supervised thermal network, leverages mutual information to transfer knowledge from the visible to the thermal spectrum for robust object detection.

b) Semi-supervised DA (Semi-DA) aims to enhance AD generalization by utilizing unlabeled target samples alongside limited labeled target data [114]. For example, SSDA3D [86] boosts 3D object detection across Waymo and nuScenes, and ECACL [115] provides a unified framework integrating categorical alignment, data augmentation, and consistency learning.

c) Unsupervised DA (UDA) tackles target domain learning using labeled source data and unlabeled target data [116] via adversarial training, self-learning, or style transfer. ST3D++ [85] applies a self-training pipeline for 3D object detection from point clouds, while SF-UDA3D [117] introduces the first source-free UDA method for radar-based 3D detection.

Despite these advancements, existing DA works face limitations [118]. Notably, performance in the source do-

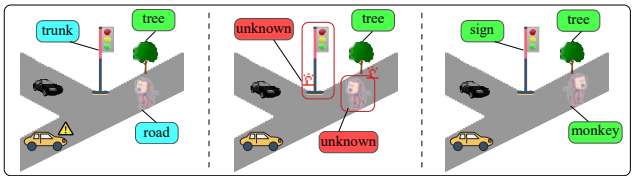


Fig. 7: Illustration of OWL-empowered AVs shows how they learn from and adapt their decision-making processes to dynamic real-world environments by integrating unknown rejection, novel class discovery, and continuous learning.

main often degrades after adaptation [119]. Furthermore, the increased model uncertainty introduced by DA poses additional safety risks for AD systems.

Zero/Few-Shot Learning (ZSL/FSL). It enables AD models to generalize to unseen categories with minimal or no labeled samples [88]. Leveraging techniques like word embedding projection, ZSL and FSL are widely used to enhance AD robustness in data-scarce or hard-to-annotate scenarios [89]. For example, ZeroDepth [90] enables ZSL target domain perception and monocular depth estimation for AD. [91] develops a reflection representation learning module and an interchange-recomposition-coherence process to improve Retinex-based image decomposition, boosting AD robustness across zero-shot day-night domains.

While ZSL and FSL offer promising data-efficient approaches for AD generalization, they heavily depend on the quality of attribute descriptions and the capture of semantic relations, challenging the efficient exploration of new categories in AD [120]. Furthermore, these models remain sensitive to adversarial examples, and ensuring their real-time timeliness remains difficult.

Open-World Learning (OWL). It integrates unknown rejection, novel class discovery, and continuous learning, enabling AVs to adapt to dynamic real-world environments and improve driving behavior over time [92], as shown in Fig. 7. Several studies have explored this paradigm: OSDA [93] leverages a contrastive vision-language pre-trained model to improve category granularity in open scenes; [121] foucses on traffic sign detection and visual odometry to support safe generalization; ADOS [94] tackles open-set perception by introducing a novel definition of objects of interest and a segmentation-context score evaluation; and the Tianmouc vision chip [122] enables accurate, low-latency, and reliable operations under extreme open-road conditions.

While OWL shows promise in specific AD applications, current performance cannot fully eliminate potential safety hazards. are needed to fully achieve it, in which the To build an advanced OWL system for AD, future multi-dimensional efforts must focus on achieving anti-fragility rather than mere robustness [123].

C. Detection and Defence of Adversarial Attacks

While AD systems offer transformative mobility, their high vulnerability to adversarial perturbations and cyber-attacks poses significant safety and reliability challenges [124]. Implementing efficient mechanisms to counter malicious risks is therefore crucial to ensure AD robustness [125]. This subsection reviews existing literature across two core themes: adversarial detection and defence, as shown in Fig. 8.

Adversarial Detection. It evaluates AD system behaviors under various malicious inputs to identify AD robustness bottlenecks [126]. To evaluate vulnerability in diverse environments, [127] introduced scene-specific optimization to enhance attack transferability in real 3D settings, while [95] leveraged VLM-generated synthetic data to assess AV performance under weather-aliasing label shift attacks. From a multimodal perspective, [96] proposed an uncertainty-based adversarial input generation method that injects sparse adversarial points to render AVs undetectable in radar-camera fusion models. Beyond static offline evaluations, EAA [98] dynamically adjusts attack strategy in real time based on the active perception-decision-control state of the AD system, while AdvSim [128] simulates safety-critical scenarios to detect systemic vulnerabilities under environmental and adversarial stress.

Adversarial Defence. It focuses on developing active countermeasures, such as model hardening and adversarial training, to maintain system security during attacks [129], [130]. For example, SIHFR [99] provides a lightweight defence to resist adversarial patch attacks on AV manufacturing and recognition protection. To counter acoustic threats, [100] presented a multimodal defence model utilizing both voice commands and 2D images against inaudible command injections. In perception domains, [131] developed a stereo regularizer that guides models to learn implicit geometric constraints between left and right images, thereby mitigating visual adversarial vectors. Additionally, for classification sub-modules, [101] introduced a structure-based learning approach that leverages self-supervised data augmentation to alleviate attack degradation.

Real-world AD development necessitate adaptive detection and defence mechanisms that balance robust security with operational performance across diverse onboard devices [126]. Promising research directions include integrated detection-defence frameworks, alongside dynamic, uncertainty-aware, and lightweight defence algorithms. Furthermore, because empirical defenses remain susceptible to evolving adversarial variants, developing certified robustness solutions [132] is critical to fundamentally break the perpetual cycle of attacks and defenses in AD environments.

V. Exactness for AD Trustworthiness

Enhancing public trust in AD requires AVs to perform high-performance tasks with real-time precision. This section focuses on the exactness of environmental sensing,

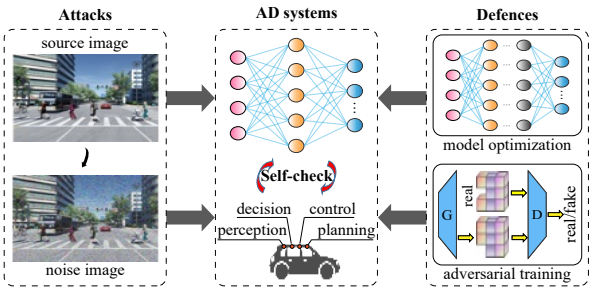


Fig. 8: Adversarial attack detection and defence in AD systems.

decision-making, and execution, achieved through both ego-AV intelligence and collaborative interconnection. We deeply explore the exact performance characterization of AD models in training, inference, and scheduling under resource constraints.

A. Real-time Precision Assurance

Real-time adaptation and precise behavioral operations are critical to the trustworthiness of AD systems [14], as users favor lower latency and higher precision. This section explores task processing performance from the perspective of real-time precision to assure exact AD.

High Precision. While DNNs readily achieve high precision in visual perception [133], as evaluated by metrics like mean average precision (mAP) and mean intersection over union (mIoU), extensive research has also focused on optimizing downstream planning and control modules. For instance, a high-precision motion controller with decoupled lateral and longitudinal error control significantly improved off-road tracking [134]. Another study achieved sub-metre (often decimetre) navigation in GPS-challenged urban environments [135]. On the KITTI dataset, a street-stereo-based 3D reconstruction method [136] yielded 94.8% pose estimation accuracy (errors $< 3°$ in orientation, < 33 cm in position). To accelerate accurate dataset labeling, DeepBbox [137] uses tight pre-labels for ground truth generation; its application on Cityscapes increased target edge precision by 37.7% with an error under 1%.

Real-time. Real-time assurance is equally vital for computing-sensitive tasks [138]. Techniques like convolution decomposition and channel shuffling are widely used for acceleration. For 3D point cloud detection, RT3D [139] utilized posture-sensitive feature mapping and pre-RoI pooling to reduce computational overhead to 0.09 seconds, shorter than typical LiDAR scanning cycles, while preserving accuracy. Furthermore, a meta-framework for benchmarking real-time segmentation [140] separated encoder and decoder modules to achieve approximately 16 fps on a Jetson TX2. MultiNet [141] unified classification, detection, and semantic segmentation at over 23 fps, while R-TOD [142] reduced the end-to-end latency of a YOLO v3-based detection system from 1070 ms to 261 ms.

Balance real-time execution and precision is a key challenge for AD exactness [143]. On the Argoverse-HD dataset (Tesla V100), a streaming perception system integrating data augmentation, Bag-of-Freebies, and Transformers achieved 34.6 stream AP [144]. To balance tracking precision and computational speed in AV nonlinear tracking, FI-MPC [145] introduced a rule-based convergence criterion framework; it outperformed LMPC in precision by 52.7% and reduced computation time by 76.3% compared to NMPC. High precision and low latency significantly enhance human trust in AD. However, technological advances also increase model and data complexity [146]. Providing high-performance collaborative computing via edge infrastructure (base stations, roadside units, and clouds) offers an efficient solution [147], making real-time precision guarantees in edge-empowered AD scenarios particularly essential.

B. Methods for Exact AD

The transition from ego-AV intelligence to collaborative interconnection moves AD model processing from reliance on a single entity to multi-node computational collaboration [148]. This section reviews current research on exact AD methods across four dimensions: efficient training, lightweight compression, accelerated inference, and resource scheduling. The visualization is shown in Fig. 9.

Distributed Parallel Training. As the scale of AD neural networks outpaces single-machine hardware capacity, accelerating parallel distributed training has become imperative [149]. Federated learning (FL) addresses this by enabling participants to upload parameters to a central server for aggregation, reducing communication overhead and accelerating AD model training [150]. For instance, a FL framework with a dynamic federated proximal algorithm optimized autonomous controllers for connected AVs, achieving a 40% faster convergence speed and significantly reduced tracking speed error on the BDD and DACT datasets [151]. Alternatively, 3D parallelism combines data (DP), tensor (TP), and pipeline parallelism (PP) to map large-scale models to hardware by partitioning computing tasks, data, and model structures [15]. For example, GPU-accelerated data-parallel pipelines optimized the frame rate and latency of an AD perception network [152]. Furthermore, to improve the training efficiency of LLMs, [153] explored the most efficient way to combine 3D parallelism on Frontier, maximizing memory utilization across multiple GPUs and reducing communication latency via hyper-parameter search.

While FL emphasizes privacy and distributed collaboration, 3D parallelism focuses on hardware-scaling optimization; both significantly enhance AD model performance. However, in the LLM era, distributed training for AD remains bounded by computing power guarantees and heterogeneous resource management, necessitating further breakthroughs in algorithms and system architectures.

Lightweight Model Compression. Deploying large-scale AD models, especially generative models like LLMs and VLMs, onto resource-constrained AVs is challenging.

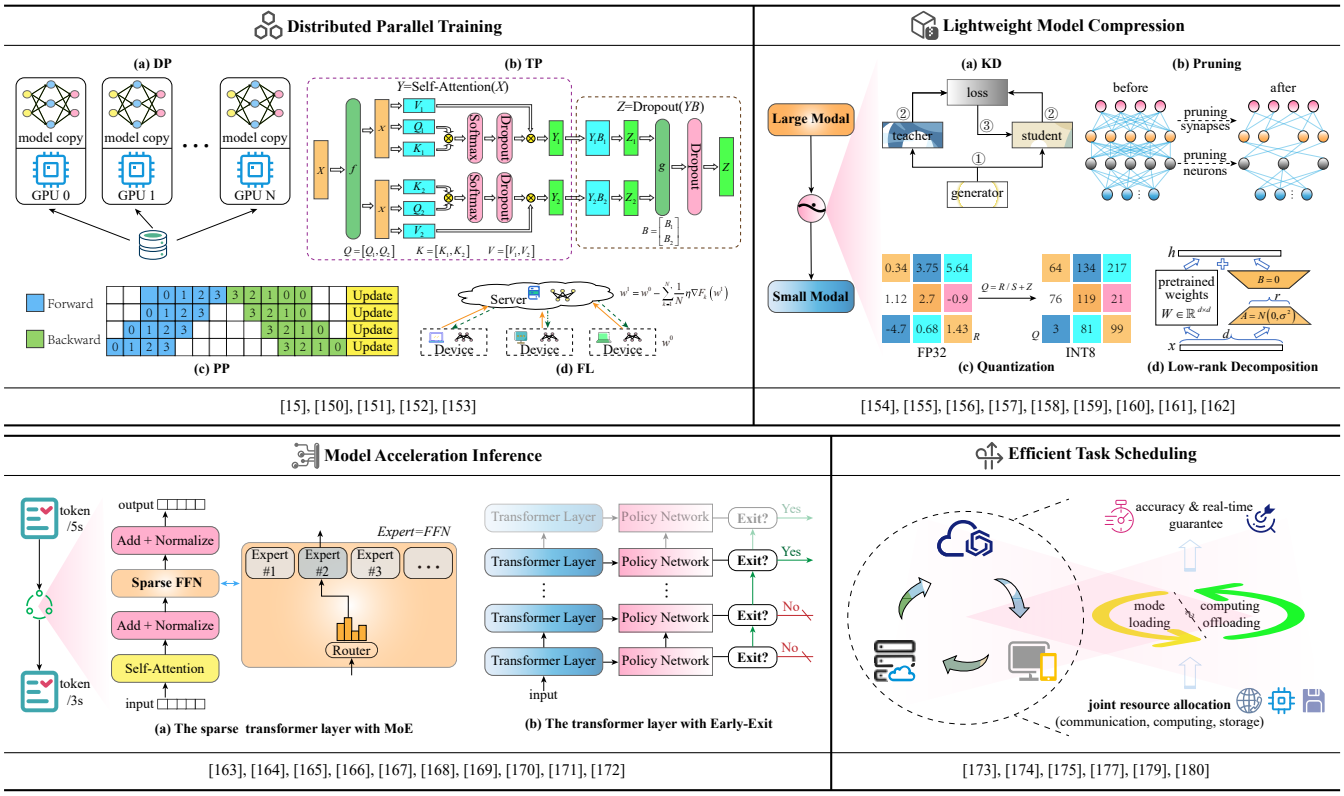


Fig. 9: Illustrative overview of approaches for achieving exact AD, including distributed model training, compression, inference optimizing, and scheduling.

Model compression addresses this by optimizing model weights or network architectures [154]. Prominent techniques include knowledge distillation (KD) based on teacher-student architectures [155], pruning by removing redundant parameters [156], quantization by converting high-precision parameters into low-precision counterparts [157], and low-rank decomposition factorizing weight matrices into low-dimensional representations [158]. For instance, itKD successfully compresses point cloud-based 3D object detection models [159]. CrossPrune, a multimodal collaborative pruning framework, efficiently removes redundant information from Camera-LiDAR features; evaluation on the KITTI and nuScenes datasets demonstrates a 9.6% relative improvement in mAP at 89.8% pruning sparsity [160]. To support AIGC-enabled AD, recent studies have integrated multiple techniques [161]. For example, LQ-LoRA decomposes pre-trained matrices into a memory-efficient quantized component and a high-precision low-rank component [162]. By updating only the low-rank component during fine-tuning, LQ-LoRA enables memory-efficient adaptation of LLMs for downstream AD tasks.

While compressed AD models accelerate task processing, they can be highly sensitive to minor perturbations, causing performance fluctuations. Therefore, further optimization is required to enhance the stability of compressed AD models while maintaining efficiency.

Model Acceleration Inference. The execution speed of deployed AD models is often constrained by onboard computing capacity, impacting latency-sensitive AD tasks. Model inference acceleration mitigates this by enabling faster task execution through model decomposition or adaptive reasoning [163], particularly within the Transformer architecture [164]. Extensively studied techniques include the gated mixture-of-experts (MoE) [165], threshold-based early-exit [166], key-value caching to reduce memory I/O [167], and speculative decoding [168]. For instance, AdaDet [169] introduces an entropy-based early-exit adaptive object detection system. Fiddler [170] optimizes the CPU-GPU dataflow, showing that running expert layers on the CPU can incur lower latency than transferring expert weights to the GPU, especially for small batches. Recent research increasingly leverages end-edge collaborative inference to enhance AD through advanced interconnection [171]. For example, ADCI [172] combines early-exit with model partitioning to enable collaborative computing offloading between devices and edge servers. Similarly, the MAB-based AdaEE [173] framework dynamically learns optimal early-exit thresholds from confidence distributions to improve cloud-edge collaborative inference.

By trading space and precision for time, these methods significantly accelerate critical perception and decision-making AD tasks. Nevertheless, practical deployment requires further consideration of power consumption, cost,

and security. Future implementations must integrate online incremental updates, adaptive mechanism adjustments, and multi-parameter optimization to satisfy the stringent real-time and safety constraints of AD.

Efficient Task Scheduling. The inherent uncertainty of complex traffic environments mandates effective task scheduling in AD systems to ensure real-time safety responses [174]. To enhance the operational efficiency of AD modules like perception, decision-making, and control, extensive research focuses on adaptively managing computing tasks across multi-dimensional resources (e.g., computing, storage, communication) and designing optimization mechanisms such as heuristics and game theory [175]. For example, with non-preemptive priority queuing theory, [176] developed a Spark Streaming framework to prioritize edge AD data, reducing the end-to-end delay of high-priority tasks by 46% in Spark 2.3.0. To mitigate latency variations from AV mobility, Portkey [177], a distributed KV storage system, treats mobility as a first-class primitive to prioritize rapid, albeit potentially suboptimal, content access. Additionally, MDTS [178] improves parallel scheduling for AD tasks; compared to particle swarm optimization, it significantly optimizes task completion time, load balance, and operating costs through experience-based allocation and exact node matching. Given the rise of AIGC, research has expanded into autonomous planning and scheduling for LLMs in AD [179]. To address LLM inference offloading and resource allocation in cloud-edge environments, an active reasoning-based method with incentive guidance was introduced [180].

While existing studies optimize scheduling via soft operations like cache placement and offloading, the heavy heterogeneity of underlying AD hardware still challenges scheduler efficacy [181]. Therefore, developing scalable scheduling frameworks with fault tolerance, priority management, and dynamic adaptation remains crucial, requiring deeper integration with the onboard hardware stack.

VI. Challenges and Opportunities

While shifting from ego-AV to multi-AV collaboration is essential for AD evolution, ensuring system trustworthiness remains critical. Fig. 10 illustrates a multi-scale, full-stack framework that connects vehicle-level micro-intelligence and network-level macro-performance, establishing a trustworthy evolutionary path for complex traffic environments.

A. Micro-scale Trustworthiness: Ego-AV Intelligence

Ensuring operational safety of ego-AVs requires confronting the inherent vulnerabilities of AI regarding transparency and resilience. These challenges crystallize into three core bottlenecks that necessitate a paradigm shift in system design.

Knowledge Representation and Generative Interpretability. While state-of-the-art vision-language-action (VLA) models excel in common-sense reasoning, their inherent opacity precludes the rigorous causal transparency required for safety-critical AD [182]. Transitioning to trustworthy AI necessitates a deeper integration of data-driven perception and logic-driven reasoning. To this end, neuro-symbolic computing for embodied intelligence offers a promising pathway [183]. By embedding explicit, structured knowledge (e.g., traffic regulations and physical priors) as high-confidence constraints, this hybrid architecture facilitates differentiable semantic alignment through multimodal chain-of-thought (CoT) reasoning [184]. In this framework, neural networks manage high-dimensional perception, while symbolic logic serves as a safety guardrail. This synergy mitigates the opacity, ensuring decision traceability and behavioral auditability.

Generalization-Robustness Trade-off in Open-World Perception. In open-world deployment, ego-AVs face a fundamental tension between broad environmental adaptability and rigorous security. Conventional defences often compromise accuracy in long-tail scenarios, while DA can introduce latent vulnerabilities to physical adversarial perturbations [185]. Current evaluation methods, largely confined to static digital domains, fail to capture the complexities of real-world physical dynamics [186]. To break this impasse, future research should pivot towards intrinsic feature-layer defence and verifiable optimization. By utilizing generative foundation models to simulate dynamic, physically consistent environments, we can proactively map safety boundaries via reverse-probing [187]. Furthermore, integrating provable safety bounds with geometric constraints enables perception systems to achieve generalized immunity against evolving, unknown physical threats.

Computing Resource Explosion vs. On-Board Hardware Constraints. Despite their unprecedented performance, multimodal foundation models achieve general intelligence through exponential parameter scaling, yet deploying such computationally demanding algorithms in ego-AVs is severely hindered by stringent physical constraints in power, thermal dissipation, and real-time response [188]. Consequently, advancing ego-AV intelligence necessitates a paradigm shift from traditional von Neumann architecture towards lightweight, edge-native, and neuromorphic computing. One promising trajectory lies in pairing brain-inspired spiking neural networks (SNNs) with AD [189]. By leveraging asynchronous, event-driven processing, this approach achieves minimal latency and microwatt-level power consumption, effectively eliminating the computational redundancy inherent in continuous forward propagation. Harnessing these spike-timing characteristics to construct adaptive control loops ensures real-time, trustworthy perception within the stringent constraints of edge hardware.

B. Macro-scale Trustworthiness: ITS

When AVs scale from isolated environments to complex urban networks, ego-AV optimization can inadvertently

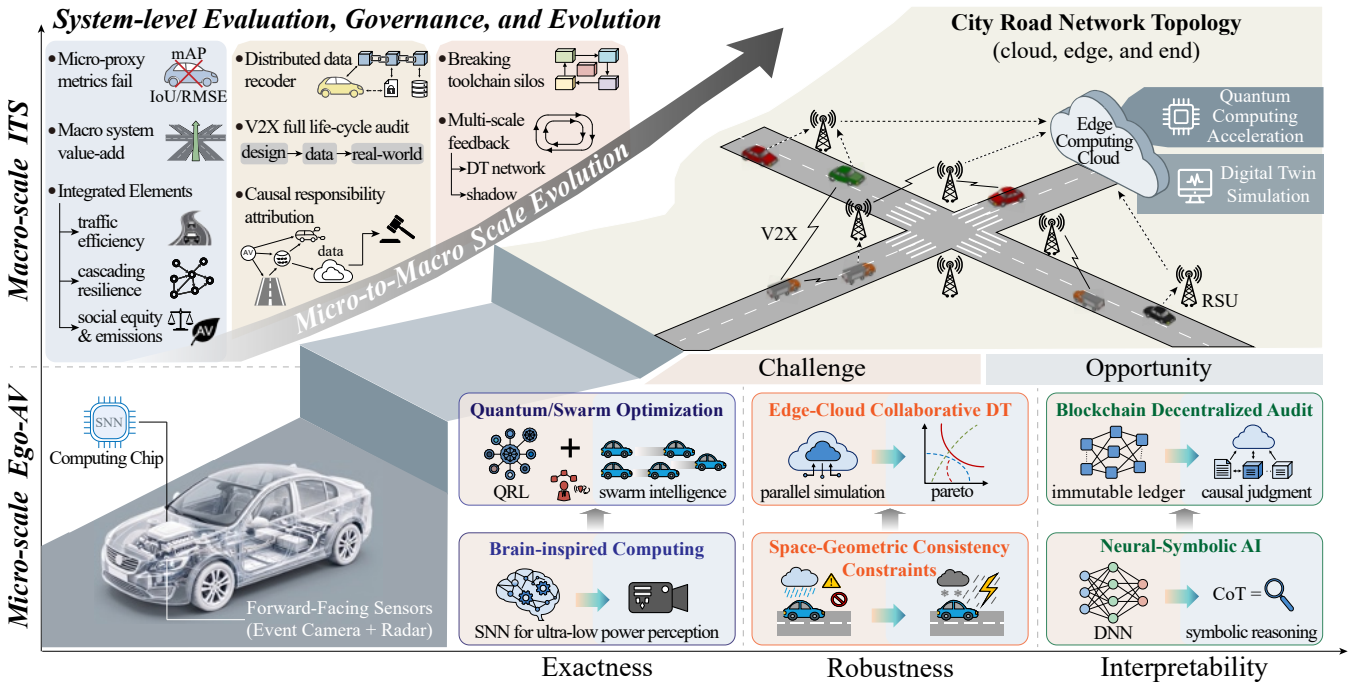


Fig. 10: A multi-scale trustworthiness framework for AI-based AD: From ego-AV intelligence to ITS. The framework maps structural bottlenecks to advanced computing paradigms and strategic solutions across the three pillars of trustworthiness: accuracy, robustness, and interpretability.

compromise global traffic efficiency. This tension necessitates a shift toward system-wide collaborative autonomy, which faces three primary challenges.

Collaborative Perception and Distributed Situational Awareness. Although ego-AV intelligence is constrained by physical line-of-sight and vulnerable to adverse weather, V2X-enabled collaborative networks effectively mitigate perception blind spots [190]. However, multi-node data synthesis poses significant challenges, notably spatiotemporal asynchrony from heterogeneous sampling, bandwidth-latency trade-offs, and network-level Byzantine faults. Recent efforts have shifted toward lightweight feature representation and adaptive alignment to resolve these bottlenecks [191]. Specifically, roadside unit (RSU)-enabled vehicle-to-infrastructure (V2I) frameworks now allow for compressing tracking data into low-bandwidth feature vectors, extending the predictive horizon to 2-3 seconds at occluded intersections. Furthermore, in global navigation satellite system (GNSS)-challenged urban environments, cooperative factor graph optimization, leveraging vehicle-to-vehicle (V2V) relative ranging, enables centimetre-level decentralized localization and high-density platoon control [192].

Multi-Agent Game Dynamics in Mixed Traffic Flows. The long-term coexistence of human-driven, connected, and AVs creates highly uncertain, multimodal traffic environments. Ego-AVs making isolated decisions often struggle with collective interactions, leading to conservative bottlenecks and stop-and-go waves. While multi-agent cooperative control offers a solution, it constitutes a non-convex, NP-hard global game; centralized computing suffers from combinatorial explosion, whereas decentralized edge nodes are limited by bandwidth and power [193]. To address this, we favour a macro-micro synergetic paradigm. Macroscopically, cloud-based quantum computing can formulate the multi-agent traffic coordination as a quadratic unconstrained binary optimization (QUBO) problem. By leveraging algorithms such as quantum annealing (QA) or quantum approximate optimization algorithm (QAOA), it extracts quantum parallelism to rapidly solve massive game spaces, facilitating a 'cloud-heavy, vehicle-light' deployment via reinforcement learning and KD [194]. Microscopically, biomimetic swarm intelligence over V2X networks (e.g., decentralized control mechanisms inspired by artificial bee colonies or bird flocking rules) enables self-organizing emergence (e.g., tight platooning and conflict-free coordination), effectively alleviating individual computational and communication burdens [195].

Edge-Cloud Digital Twins for Network-Wide Optimization. Rising AV penetration necessitates a shift from egocentric safety to network-wide coordination. Traditional decentralized ego-AVs routing or localized V2X triggers often lead to selfish behaviors, converging toward a sub-optimal Nash equilibrium and risking systemic gridlock [196]. Achieving global optimality remains hindered by spatiotemporal mismatches: millisecond-level trajectory planning versus minute-level macroscopic regulation—a latency-optimality tradeoff that centralized architectures fail to reconcile [197]. To bridge this, we embrace a 'macro-bounding, micro-adaptive' closed-loop digital twin (DT) architecture leveraging edge-cloud collaboration [198].

The cloud DT performs parallel simulations to forecast network dynamics and define macroscopic fundamental diagram boundaries. These boundaries are downlinked to dynamically constrain the solution space of distributed AV planners, ensuring millisecond-level local responsiveness while driving global convergence. For instance, during traffic incidents, the cloud predicts shockwave propagation and allocates diversion quotas, empowering edge servers to prune routes proactively and mitigate gridlock at the source.

C. System-Level Evaluation, Governance, and Evolution

Validating the systemic contribution of AI-based AD requires a fundamental paradigm shift: moving beyond isolated, algorithm-centric metrics toward holistic, system-oriented value assessments. This transition highlights three key governance bottlenecks and outlines corresponding evolutionary pathways for the ecosystem.

Cross-Scale, Multidimensional System-Level Evaluation. Current AD evaluation paradigms predominantly rely on vehicle-centric, micro-level proxy metrics, such as mAP, IoU, and root mean square error (RMSE). While essential for functional validation, these metrics fail to capture the broader impacts of large-scale AD deployment on macro-level road network operations [199]. To bridge this gap, future frameworks must adopt a cross-scale, multidimensional approach that evaluates system-level synergies [200]. Specifically, assessment must shift toward integrating vehicle-level functional safety with macro-level traffic efficiency (e.g., throughput optimization and congestion reduction), network-wide cascading resilience, full-lifecycle emissions control, data governance, and social mobility equity [201].

Distributed Governance and Full-Lifecycle Auditing. As AVs increasingly merge into vehicle-road-cloud ecosystems as mobile network nodes, the frequent exchange of multi-source, heterogeneous data introduces significant cybersecurity and privacy challenges. To address these, a blockchain-based paradigm, leveraging decentralized trust consensus and tamper-proof ledgers, offers a compelling solution [202]. This framework facilitates full-lifecycle compliance auditing across V2X collaborations, functioning as a distributed 'black box' for data accountability. Crucially, it provides a mathematical foundation for formalizing causal liability attribution among ego-AVs, roadside infrastructure, and cloud service providers, thereby establishing a transparent and trustworthy framework for modern transportation governance [203].

Multi-Level Cyber-Physical Testing Ecosystems. Deploying highly reliable AD systems necessitates integrated, closed-loop testing platforms capable of handling multi-scale feedback. Recent empirical studies [204] have highlighted that despite various emerging technical solutions, there remains a persistent gap between academic algorithmic validation and the practical testing needs of industry practitioners. Therefore, a significant opportunity lies in breaking down existing toolchain silos to fully integrate physical closed-course testing, microscopic traffic simulation, regional dynamic DTs, and real-world 'shadow mode' data [205]. By continuously auditing and benchmarking the system-level impacts of AI-based AD within this multi-level matrix, we can effectively bridge the critical gap between algorithmic validation and trustworthy commercial deployment.

Synthesis: Balancing Trustworthiness for Real-World Applicability. Ultimately, translating AI-based AD from academic benchmarks to real-world commercial deployment requires navigating the inherent trade-offs among exactness, robustness, and interpretability. As synthesized throughout this section, practical paradigms offer the essential bridges to navigate these trade-offs. For instance, neuro-symbolic architectures combine the robust perception capabilities of deep learning with the exactness and interpretability of symbolic logic. By embedding explicit traffic regulations and safety constraints into the learning process, such architectures ensure that the AD system not only adapts to complex, dynamic environments (robustness) but also provides human-readable justifications for its critical decisions (interpretability), thereby paving the way for real-world certification and deployment.

VII. Conclusion

This research provides a comprehensive review of trustworthy AD across the core domains of perception, decision-making, planning, and control. We first examine how traditional AI and LLM paradigms enhance the interpretability of AD behavior. We then address strategies for improving AD robustness across diverse scenarios, specifically focusing on missing data completion, model generalization, as well as attack detection and defense mechanisms. To further optimize operational exactness, we explore solutions in distributed training, model compression, inference acceleration, and task scheduling. By synthesizing these, we propose a multi-scale trustworthy evolution framework that bridges micro-scale ego-AV intelligence with macro-scale ITS to foster reliable AD; meanwhile, we discuss the remaining challenges and opportunities to realize this vision.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62172438, Shenzhen Science and Technology Program JCYJ20220818103200002 and the Natural Science Foundation of Guangdong Province under Grant No. 2024A1515011155.

References

[1] M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, Autonomous driving: technical, legal and social aspects. Springer Nature, 2016.

[2] Precedence Research. (2026) Autonomous vehicle market size, share and trends 2026 to 2035. [Online]. Available: https://www.precedenceresearch.com/autonomous-vehicle-market

[3] N. E. Boudette, "Tesla's self-driving system cleared in deadly crash," The New York Times, 2017.

[4] R. E. Stern, S. Cui, M. L. Delle Monache, R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, R. Haulcy, H. Pohlmann, F. Wu et al., "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," Transportation research part C: emerging technologies, vol. 89, pp. 205–221, 2018.

[5] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," Transportation research part C: emerging technologies, vol. 71, pp. 143–163, 2016.

[6] M. Zhu, Y. Wang, Z. Pu, J. Hu, X. Wang, and R. Ke, "Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving," Transportation Research Part C: Emerging Technologies, vol. 117, p. 102662, 2020.

[7] Y. Zheng, J. Wang, and K. Li, "Smoothing traffic flow via control of autonomous vehicles," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 3882–3896, 2020.

[8] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone et al., "Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking," Information, vol. 13, no. 8, p. 395, 2022.

[9] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards safe, explainable, and regulated autonomous driving," in Explainable Artificial Intelligence for Intelligent Transportation Systems. CRC Press, 2021, pp. 32–52.

[10] Y. Hua, D. Zhang, and S. Ge, "Research progress on the interpretability of deep learning models," Journal of Cyber Security, vol. 5, no. 3, 2020.

[11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," AI magazine, vol. 38, no. 3, pp. 50–57, 2017.

[12] S. Nageshrao, Y. Rahman, V. Ivanovic, M. Jankovic, E. Tseng, M. Hafner, and D. Filev, "Robust ai driving strategy for autonomous vehicles," in AI-enabled Technologies for Autonomous and Connected Vehicles. Springer, 2022, pp. 161–212.

[13] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 444–453.

[14] K. Chen, R. Oldja, N. Smolyanskiy, S. Birchfield, A. Popov, D. Wehr, I. Eden, and J. Pehserl, "Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 2288–2294.

[15] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti et al., "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model," arXiv preprint arXiv:2201.11990, 2022.

[16] C. Chen, Z. Liu, W. Jiang, G. S. Qi, and K.-Y. Lam, "Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations," arXiv preprint arXiv:2408.12935, 2024.

[17] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–20, 2024.

[18] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.

[19] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7644–7652.

[20] J. Zhang, F. Li, F. Ye, and H. Wu, "Autonomous unknown-application filtering and labeling for dl-based traffic classifier update," in IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2020, pp. 397–405.

[21] Y. Lv, "Artificial intelligence-generated content in intelligent transportation systems: Learning to copy, change, and create![editor's column]," IEEE Intelligent Transportation Systems Magazine, vol. 15, no. 5, pp. 2–3, 2023.

[22] M. Xu, D. Niyato, J. Chen, H. Zhang, J. Kang, Z. Xiong, S. Mao, and Z. Han, "Generative ai-empowered simulation for autonomous driving in vehicular mixed reality metaverses," arXiv preprint arXiv:2302.08418, 2023.

[23] C. Chen, J. Fu, and L. Lyu, "A pathway towards responsible ai generated content," arXiv preprint arXiv:2303.01325, 2023.

[24] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," arXiv preprint arXiv:2311.17911, 2023.

[25] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2942–2950.

[26] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, "Visualbackprop: Efficient visualization of cnns for autonomous driving," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 4701–4708.

[27] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, and V. Fischer, "Grid saliency for context explanations of semantic segmentation," Advances in neural information processing systems, vol. 32, 2019.

[28] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 340–341.

[29] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9523–9532.

[30] W. Bao, Q. Yu, and Y. Kong, "Drive: Deep reinforced accident anticipation with visual explanation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7619–7628.

[31] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15 793–15 803.

[32] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 13 731–13 737.

[33] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2902–2911.

[34] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5068–5078, 2021.

[35] A. Stocco, P. J. Nunes, M. d'Amorim, and P. Tonella, "Thirdeye: Attention maps for safe autonomous driving systems," in Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, 2022, pp. 1–12.

[36] K. Renz, K. Chitta, O.-B. Mercea, A. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," arXiv preprint arXiv:2210.14222, 2022.

[37] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," in Conference and Workshop on Neural Information Processing Systems, 2023.

[38] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in International Conference on Robotics and Automation (ICRA), 2024.

[39] W. Ding, I. Alrashdi, H. Hawash, and M. Abdel-Basset, "Deepsecdrive: An explainable deep learning framework for real-time detection of cyberattack in in-vehicle networks," Information Sciences, vol. 658, p. 120057, 2024.

[40] J. Echterhoff, A. Yan, K. Han, A. Abdelraouf, R. Gupta, and J. McAuley, "Driving through the concept gridlock: Unraveling

explainability bottlenecks in automated driving," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7346–7355.
[41] S. Park, M. Lee, J. Kang, H. Choi, Y. Park, J. Cho, A. Lee, and D. Kim, "Vlaad: Vision and language assistant for autonomous driving," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 980–987.
[42] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," in Robotics: Science and Systems, 2024.
[43] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," in International Conference on Learning Representations, 2024.
[44] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 10 142–10 162, 2021.
[45] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.
[46] R. Nahata, D. Omeiza, R. Howard, and L. Kunze, "Assessing and explaining collision risk in dynamic environments for autonomous driving safety," in 2021 IEEE international intelligent transportation systems conference (ITSC). IEEE, 2021, pp. 223–230.
[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[48] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018, pp. 839–847.
[49] A. Marathe, P. Jain, R. Walambe, and K. Kotecha, "Restorex-ai: A contrastive approach towards guiding image restoration via explainable ai systems," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3030–3039.
[50] Y. M. Yousif and J. P. Müller, "Generating explanatory saliency maps for mixed traffic flow using a behaviour cloning model," in International Workshop on Multi-Agent Systems and Agent-Based Simulation. Springer, 2022, pp. 107–120.
[51] S. Mohseni, A. Jagadeesh, and Z. Wang, "Predicting model failure using saliency maps in autonomous driving systems," arXiv preprint arXiv:1905.07679, 2019.
[52] A. Samadi, A. Shirian, K. Koufos, K. Debattista, and M. Dianati, "Safe: Saliency-aware counterfactual explanations for dnn-based automated driving systems," in 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2023, pp. 5655–5662.
[53] X. Lu, A. Tolmachev, T. Yamamoto, K. Takeuchi, S. Okajima, T. Takebayashi, K. Maruhashi, and H. Kashima, "Crowdsourcing evaluation of saliency-based xai methods," in Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V 21. Springer, 2021, pp. 431–446.
[54] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 563–578.
[55] S. Jain and B. C. Wallace, "Attention is not explanation," arXiv preprint arXiv:1902.10186, 2019.
[56] H. Wang, H. Gao, S. Yuan, H. Zhao, K. Wang, X. Wang, K. Li, and D. Li, "Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning," IEEE Transactions on Vehicular Technology, vol. 70, no. 9, pp. 8707–8719, 2021.
[57] M. Itkina and M. Kochenderfer, "Interpretable self-aware neural networks for robust trajectory prediction," in Conference on Robot Learning. PMLR, 2023, pp. 606–617.
[58] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in Conference on Robot Learning. PMLR, 2023, pp. 726–737.
[59] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," arXiv preprint arXiv:2312.03661, 2023.
[60] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7699–7707.
[61] T. You and B. Han, "Traffic accident benchmark for causality recognition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer, 2020, pp. 540–556.
[62] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," arXiv preprint arXiv:1909.10838, 2019.
[63] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? a new dataset for anomaly detection in driving videos," arXiv preprint arXiv:2004.03044, 2020.
[64] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in Conference on Robot Learning. PMLR, 2021, pp. 409–418.
[65] Y. Shen, S. Jiang, Y. Chen, and K. D. Campbell, "To explain or not to explain: A study on the necessity of explanations for autonomous vehicles," arXiv preprint arXiv:2006.11684, 2020.
[66] Z. Yang, X. Jia, H. Li, and J. Yan, "Llm4drive: A survey of large language models for autonomous driving," arXiv e-prints, pp. arXiv–2311, 2023.
[67] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu et al., "Palm-e: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.
[68] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," arXiv preprint arXiv:2310.01412, 2023.
[69] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 902–909.
[70] C. Cui, X. Cao, Y. Ma, W. Ye, and Z. Wang, "Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles," IEEE Intelligent Transportation Systems Magazine, vol. 16, no. 4, pp. 81–94, 2024.
[71] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 910–919.
[72] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," arXiv preprint arXiv:2311.10813, 2023.
[73] H. Liao, Z. Li, H. Shen, W. Zeng, D. Liao, G. Li, and C. Xu, "Bat: Behavior-aware human-like trajectory prediction for autonomous driving," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 9, 2024, pp. 10 332–10 340.
[74] L. Zhao, X. Fan, A. Hawbani, L. Xu, K. Yu, Z. Liu, and O. Alfarraj, "Generative abnormal data detection for enhancing cellular vehicle-to-everything based road safety," IEEE Transactions on Green Communications and Networking, pp. 1–1, 2024.
[75] L. Zhao, H. Chai, Y. Han, K. Yu, and S. Mumtaz, "A collaborative v2x data correction method for road safety," IEEE Transactions on Reliability, vol. 71, no. 2, pp. 951–962, 2022.
[76] K. Gao, B. Wang, L. Xiao, and G. Mei, "Incomplete road information imputation using parallel interpolation to enhance the safety of autonomous driving," IEEE Access, vol. 8, pp. 25 420–25 430, 2020.
[77] S. Chen, Y. Ma, Y. Qiao, and Y. Wang, "M-bev: Masked bev perception for robust autonomous driving," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 2, 2024, pp. 1183–1191.

[78] C. Zhang, Z. Huang, M. H. Ang, and D. Rus, "Lidar missing measurement detection for autonomous driving in rain," in 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023, pp. 7393–7399.

[79] Y. Li, L. Yu, L. Xing, and F. Liu, "Analysis of influencing factors of lane change prediction with data missing," IEEE Transactions on Intelligent Vehicles, 2024.

[80] R. C. Pereira, M. S. Santos, P. P. Rodrigues, and P. H. Abreu, "Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes," Journal of Artificial Intelligence Research, vol. 69, pp. 1255–1285, 2020.

[81] T. Zheng, A. Li, Z. Chen, H. Wang, and J. Luo, "Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving," in Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023, pp. 1–15.

[82] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14 943–14 952.

[83] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for 3d detection and map segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8721–8731.

[84] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, 2021, pp. 2302–2310.

[85] J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, "St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 5, pp. 6354–6371, 2022.

[86] Y. Wang, J. Yin, W. Li, P. Frossard, R. Yang, and J. Shen, "Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 3, 2023, pp. 2707–2715.

[87] D. Kothandaraman, R. Chandra, and D. Manocha, "Ss-sfda: Self-supervised source-free domain adaptation for road segmentation in hazardous environments," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3049–3059.

[88] Y. Lu, Q. Jiang, R. Chen, Y. Hou, X. Zhu, and Y. Ma, "See more and know more: Zero-shot point cloud segmentation via multi-modal visual data," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21 674–21 684.

[89] J. Mei, J. Zhou, and Y. Hu, "Few-shot 3d lidar semantic segmentation for autonomous driving," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 9324–9330.

[90] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambruș, and A. Gaidon, "Towards zero-shot scale-aware monocular depth estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9233–9243.

[91] Z. Du, M. Shi, and J. Deng, "Boosting object detection with zero-shot day-night domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12 666–12 676.

[92] Z. Ma, Y. Yang, G. Wang, X. Xu, H. T. Shen, and M. Zhang, "Rethinking open-world object detection in autonomous driving scenarios," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1279–1288.

[93] Z. Ma, Z. Zheng, J. Wei, X. Wei, Y. Yang, and H. T. Shen, "Open-scenario domain adaptive object detection in autonomous driving," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 8453–8462.

[94] C. Bunel, M. Gueriau, A. Daoud, S. Ainouz, and G. Gasso, "Addressing open-set object detection for autonomous driving perception: A focus on road objects."

[95] A. Marathe and S. Prabhu, "t-rain: Robust generalization under weather-aliasing label shift attacks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5844–5853.

[96] D. Unal, F. O. Catak, M. T. Houkan, M. Mudassir, and M. Hammoudeh, "Towards robust autonomous driving systems through adversarial test set generation," ISA transactions, vol. 132, pp. 69–79, 2023.

[97] B. Yang, X. Ji, Z. Jin, Y. Cheng, and W. Xu, "Exploring adversarial robustness of lidar-camera fusion model in autonomous driving," in 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2). IEEE, 2023, pp. 3634–3639.

[98] Y. Sun, Y. Huang, and X. Wei, "Embodied adversarial attack: A dynamic robust physical attack in autonomous driving," arXiv preprint arXiv:2312.09554, 2023.

[99] A. J. Siddiqui and A. Boukerche, "A novel lightweight defense method against adversarial patches-based attacks on automated vehicle make and model recognition systems," Journal of Network and Systems Management, vol. 29, no. 4, p. 41, 2021.

[100] J. Guan, X. Zheng, C. Wang, Y. Zhou, and A. Jolfaei, "Robust sensor fusion algorithms against voice command attacks in autonomous vehicles," in 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2021, pp. 895–902.

[101] M. K. Sharma, R. Choudhuri, M. Dixit, M. Sarkar, and B. Dittakavi, "Structure-based learning for robust defense against adversarial attacks in autonomous driving agents," in International Conference on Computer Vision and Image Processing. Springer, 2022, pp. 474–489.

[102] J. Zou, T. Huang, G. Yang, Z. Guo, and W. Zuo, "Unim ˆ2 ae: Multi-modal masked autoencoders with unified 3d representation for 3d perception in autonomous driving," arXiv preprint arXiv:2308.10421, 2023.

[103] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani, "Modality-agnostic learning for radar-lidar fusion in vehicle detection," in proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 918–927.

[104] F. Barbato, E. Camuffo, S. Milani, and P. Zanuttigh, "Continual road-scene semantic segmentation via feature-aligned symmetric multi-modal network," arXiv preprint arXiv:2308.04702, 2023.

[105] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[106] S. Akhauri, L. Zheng, T. Goldstein, and M. Lin, "Improving generalization of transfer learning across domains using spatio-temporal features in autonomous driving," arXiv preprint arXiv:2103.08116, 2021.

[107] V. Kilic, D. Hegde, V. Sindagi, A. B. Cooper, M. A. Foster, and V. M. Patel, "Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection," arXiv preprint arXiv:2107.07004, 2021.

[108] W. Wu, X. Deng, P. Jiang, S. Wan, and Y. Guo, "Crossfuser: multi-modal feature fusion for end-to-end autonomous driving under unseen weather conditions," IEEE Transactions on Intelligent Transportation Systems, 2023.

[109] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer, 2020, pp. 763–780.

[110] H. Gupta, O. Kotlyar, H. Andreasson, and A. J. Lilienthal, "Robust object detection in challenging weather conditions," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2024, pp. 7508–7517.

[111] T.-H. Le, S.-C. Huang, Q.-V. Hoang, Z. Lokaj, and Z. Lu, "Amalgamating knowledge for object detection in rainy weather conditions," ACM Transactions on Intelligent Systems and Technology, vol. 16, no. 2, pp. 1–20, 2025.

[112] E. S. Lee, J. Kim, and Y. M. Kim, "Self-supervised domain adaptation for visual navigation with global map consistency," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 1707–1716.

[113] F. Munir, S. Azam, and M. Jeon, "Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving," in 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2021, pp. 206–213.

[114] Y. S. Tang and G. H. Lee, "Transferable semi-supervised 3d object detection from rgb-d data," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1931–1940.

[115] K. Li, C. Liu, H. Zhao, Y. Zhang, and Y. Fu, "Ecacl: A holistic framework for semi-supervised domain adaptation," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8578–8587.

[116] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: a review," Technologies, vol. 8, no. 2, p. 35, 2020.

[117] C. Saltori, S. Lathuiliére, N. Sebe, E. Ricci, and F. Galasso, "Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection," in 2020 International Conference on 3D Vision (3DV). IEEE, 2020, pp. 771–780.

[118] Y. Mitsuzumi, G. Irie, D. Ikami, and T. Shibata, "Generalized domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1084–1093.

[119] S. Yang, Y. Wang, J. Van De Weijer, L. Herranz, and S. Jui, "Generalized source-free domain adaptation," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8978–8987.

[120] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 69–77.

[121] S. Li, X. Wu, Y. Cao, and H. Zha, "Generalizing to the open world: Deep visual odometry with online adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 184–13 193.

[122] Z. Yang, T. Wang, Y. Lin, Y. Chen, H. Zeng, J. Pei, J. Wang, X. Liu, Y. Zhou, J. Zhang et al., "A vision chip with complementary pathways for open-world sensing," Nature, vol. 629, no. 8014, pp. 1027–1033, 2024.

[123] M. Kejriwal, E. Kildebeck, R. Steininger, and A. Shrivastava, "Challenges, evaluation and opportunities for open-world learning," Nature Machine Intelligence, pp. 1–9, 2024.

[124] H. Ren, T. Huang, and H. Yan, "Adversarial examples: attacks and defenses in the physical world," International Journal of Machine Learning and Cybernetics, vol. 12, no. 11, pp. 3325–3336, 2021.

[125] X. Ma, K. Driggs-Campbell, and M. J. Kochenderfer, "Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning," in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1665–1671.

[126] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2280–2289.

[127] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, and G. Buttazzo, "On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving," IEEE Transactions on Neural Networks and Learning Systems, 2023.

[128] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9909–9918.

[129] K. Y. Mahima, M. Ayoob, and G. Poravi, "Adversarial attacks and defense technologies on autonomous vehicles: A review." Appl. Comput. Syst., vol. 26, no. 2, pp. 96–106, 2021.

[130] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," arXiv preprint arXiv:1902.06705, 2019.

[131] Q. Sun, A. A. Rao, X. Yao, B. Yu, and S. Hu, "Counteracting adversarial attacks in autonomous driving," in Proceedings of the 39th International Conference on Computer-Aided Design, 2020, pp. 1–7.

[132] L. Li, T. Xie, and B. Li, "Sok: Certified robustness for deep neural networks," in 2023 IEEE symposium on security and privacy (SP). IEEE, 2023, pp. 1289–1310.

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

[134] Z. Sun, Z. Huang, Q. Zhu, X. Li, and D. Liu, "High-precision motion control method and practice for autonomous driving in complex off-road environments," in 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2016, pp. 767–773.

[135] S. Zhao, Y. Chen, and J. A. Farrell, "High-precision vehicle navigation in urban environments using an mem's imu and single-frequency gps receiver," IEEE transactions on intelligent transportation systems, vol. 17, no. 10, pp. 2854–2867, 2016.

[136] M. Coenen, F. Rottensteiner, and C. Heipke, "Precise vehicle reconstruction for autonomous driving applications," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; IV-2/W5, vol. 4, no. 2/W5, pp. 21–28, 2019.

[137] G. Rathore, W.-Y. Lin, and J. E. Kim, "Deepbbox: accelerating precise ground truth generation for autonomous driving datasets," in 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 1871–1876.

[138] B. Ravi Kiran, L. Roldao, B. Irastorza, R. Verastegui, S. Suss, S. Yogamani, V. Talpaert, A. Lepoutre, and G. Trehard, "Real-time dynamic object detection for autonomous driving using prior 3d-maps," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

[139] Y. Zeng, Y. Hu, S. Liu, J. Ye, Y. Han, X. Li, and N. Sun, "Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving," IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3434–3440, 2018.

[140] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, "A comparative study of real-time semantic segmentation for autonomous driving," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 587–597.

[141] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in 2018 IEEE intelligent vehicles symposium (IV). IEEE, 2018, pp. 1013–1020.

[142] W. Jang, H. Jeong, K. Kang, N. Dutt, and J.-C. Kim, "R-tod: Real-time object detector with minimized end-to-end delay for autonomous driving," in 2020 IEEE Real-Time Systems Symposium (RTSS). IEEE, 2020, pp. 191–204.

[143] S. Miraliev, S. Abdigapporov, V. Kakani, and H. Kim, "Real-time memory efficient multitask learning model for autonomous driving," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 247–258, 2023.

[144] Y. Gu, Q. Wang, and X. Qin, "Real-time streaming perception system for autonomous driving," in 2021 China Automation Congress (CAC). IEEE, 2021, pp. 5239–5244.

[145] D. Meng, H. Chu, M. Tian, B. Gao, and H. Chen, "Real-time high-precision nonlinear tracking control of autonomous vehicles using fast iterative model predictive control," IEEE Transactions on Intelligent Vehicles, 2024.

[146] C. Katrakazas, M. Quddus, W.-H. Chen, and L. Deka, "Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions," Transportation Research Part C: Emerging Technologies, vol. 60, pp. 416–442, 2015.

[147] S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "Vips: Real-time perception fusion for infrastructure-assisted autonomous driving," in Proceedings of the 28th annual international conference on mobile computing and networking, 2022, pp. 133–146.

[148] L. Zhao, Y. Han, A. Hawbani, S. Wan, Z. Guo, and M. Guizani, "Media: An incremental dnn based computation offloading for collaborative cloud-edge computing," IEEE Transactions on Network Science and Engineering, vol. 11, no. 2, pp. 1986–1998, 2024.

[149] H. Zhang, J. Bosch, and H. H. Olsson, "End-to-end federated learning for autonomous driving vehicles," in 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021, pp. 1–8.

[150] A. Nguyen, T. Do, M. Tran, B. X. Nguyen, C. Duong, T. Phan, E. Tjiputra, and Q. D. Tran, "Deep federated learning for autonomous driving," in 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022, pp. 1824–1830.

[151] T. Zeng, O. Semiari, M. Chen, W. Saad, and M. Bennis, "Federated learning on the road autonomous controller design for connected and autonomous vehicles," IEEE Transactions on Wireless Communications, vol. 21, no. 12, pp. 10 407–10 423, 2022.

[152] S. Ahn, S. Kim, H. Kang, and J.-C. Kim, "Data-parallel real-time perception system with partial gpu acceleration for autonomous driving," in Proc. 8th Machine Learning for Autonomous Driving Symposium (ML4AD), 2023.

[153] S. Dash, I. R. Lyngaas, J. Yin, X. Wang, R. Egele, J. A. Ellis, M. Maiterth, G. Cong, F. Wang, and P. Balaprakash, "Optimizing distributed training on frontier for large language models," in ISC High Performance 2024 Research Paper Proceedings (39th International Conference). Prometeus GmbH, 2024, pp. 1–11.

[154] H. Liu, Y. He, F. R. Yu, and J. James, "Flexi-compression: a flexible model compression method for autonomous driving," in Proceedings of the 11th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications, 2021, pp. 19–26.

[155] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8637–8646.

[156] C. Yang, P. Zhao, Y. Li, W. Niu, J. Guan, H. Tang, M. Qin, B. Ren, X. Lin, and Y. Wang, "Pruning parameterization with bi-level optimization for efficient semantic segmentation on the edge," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15 402–15 412.

[157] S. J. Rengarajan, "Mixed precision quantization for computer vision tasks in autonomous driving," 2022.

[158] L. Zhao, X. Xu, Z. Wang, Y. Zhang, B. Zhang, W. Zheng, D. Du, J. Zhou, and J. Lu, "Lowrankocc: Tensor decomposition and low-rank recovery for vision-based 3d semantic occupancy prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9806–9815.

[159] H. Cho, J. Choi, G. Baek, and W. Hwang, "itkd: Interchange transfer-based knowledge distillation for 3d object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13 540–13 549.

[160] Y. Lu, B. Jiang, N. Liu, Y. Li, J. Chen, Y. Zhang, and Z. Wan, "Crossprune: Cooperative pruning for camera–lidar fused perception models of autonomous driving," Knowledge-Based Systems, vol. 289, p. 111522, 2024.

[161] Y. Xu, L. Xie, X. Gu, X. Chen, H. Chang, H. Zhang, Z. Chen, X. Zhang, and Q. Tian, "Qa-lora: Quantization-aware low-rank adaptation of large language models," arXiv preprint arXiv:2309.14717, 2023.

[162] H. Guo, P. Greengard, E. P. Xing, and Y. Kim, "Lq-lora: Low-rank plus quantized matrix decomposition for efficient language model finetuning," arXiv preprint arXiv:2311.12023, 2023.

[163] H. Qi, F. Ren, L. Wang, P. Jiang, S. Wan, and X. Deng, "Multi-compression scale dnn inference acceleration based on cloud-edge-end collaboration," ACM Transactions on Embedded Computing Systems, vol. 23, no. 1, 2024.

[164] J. Zhong, Z. Liu, and X. Chen, "Transformer-based models and hardware acceleration analysis in autonomous driving: A survey," arXiv preprint arXiv:2304.10891, 2023.

[165] T. Chen, X. Chen, X. Du, A. Rashwan, F. Yang, H. Chen, Z. Wang, and Y. Li, "Adamv-moe: Adaptive multi-task vision mixture-of-experts," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17 346–17 357.

[166] S. Bae, J. Ko, H. Song, and S.-Y. Yun, "Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding," Conference on Empirical Methods in Natural Language Processing, 2023.

[167] W. Lee, J. Lee, J. Seo, and J. Sim, "Infinigen: Efficient generative inference of large language models with dynamic kv cache management," arXiv preprint arXiv:2406.19707, 2024.

[168] Y. Leviathan, M. Kalman, and Y. Matias, "Fast inference from transformers via speculative decoding," in International Conference on Machine Learning. PMLR, 2023, pp. 19 274–19 286.

[169] L. Yang, Z. Zheng, J. Wang, S. Song, G. Huang, and F. Li, "Adadet: An adaptive object detection system based on early-exit neural networks," IEEE Transactions on Cognitive and Developmental Systems, vol. 16, no. 1, pp. 332–345, 2023.

[170] K. Kamahori, Y. Gu, K. Zhu, and B. Kasikci, "Fiddler: Cpu-gpu orchestration for fast inference of mixture-of-experts models," arXiv preprint arXiv:2402.07033, 2024.

[171] F. Liu, J. Chen, Q. Zhang, and B. Li, "Online mec offloading for v2v networks," IEEE Transactions on Mobile Computing, vol. 22, no. 10, pp. 6097–6109, 2022.

[172] X. Zhang, Y. Teng, N. Wang, B. Sun, and G. Hu, "Accelerating deep neural network tasks through edge-device adaptive inference," in 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2023, pp. 1–6.

[173] R. G. Pacheco, M. Shifrin, R. S. Couto, D. S. Menasché, M. K. Hanawal, and M. E. M. Campista, "Adaee: Adaptive early-exit dnn inference through multi-armed bandits," in ICC 2023-IEEE International Conference on Communications. IEEE, 2023, pp. 3726–3731.

[174] J. Ma, L. Li, and C. Xu, "Autors: Environment-dependent real-time scheduling for end-to-end autonomous driving," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 12, pp. 3238–3252, 2023.

[175] C. Xu, Q. Xu, J. Wang, K. Wu, K. Lu, and C. Qiao, "Aoi-centric task scheduling for autonomous driving systems," in IEEE INFOCOM 2022-IEEE Conference on Computer Communications. IEEE, 2022, pp. 1019–1028.

[176] L. Yao, H. Zhao, J. Tang, S. Liu, and J.-L. Gaudiot, "Streaming data priority scheduling framework for autonomous driving by edge," in 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2021, pp. 37–42.

[177] J. Noor, M. Srivastava, and R. Netravali, "Portkey: Adaptive key-value placement over dynamic edge networks," in Proceedings of the ACM Symposium on Cloud Computing, 2021, pp. 197–213.

[178] Q. Qi, L. Zhang, J. Wang, H. Sun, Z. Zhuang, J. Liao, and F. R. Yu, "Scalable parallel task scheduling for autonomous driving using multi-task deep reinforcement learning," IEEE Transactions on Vehicular Technology, vol. 69, no. 11, pp. 13 861–13 874, 2020.

[179] V. Pallagani, B. C. Muppasani, K. Roy, F. Fabiano, A. Loreggia, K. Murugesan, B. Srivastava, F. Rossi, L. Horesh, and A. Sheth, "On the prospects of incorporating large language models (llms) in automated planning and scheduling (aps)," in Proceedings of the International Conference on Automated Planning and Scheduling, vol. 34, 2024, pp. 432–444.

[180] Y. He, J. Fang, F. R. Yu, and V. C. Leung, "Large language models (llms) inference offloading and resource allocation in cloud-edge computing: An active inference approach," IEEE Transactions on Mobile Computing, 2024.

[181] A. Amarnath, S. Pal, H. Kassa, A. Vega, A. Buyuktosunoglu, H. Franke, J.-D. Wellman, R. Dreslinski, and P. Bose, "Hetsched: Quality-of-mission aware scheduling for autonomous vehicle socs," arXiv preprint arXiv:2203.13396, 2022.

[182] S. Jiang, Z. Huang, K. Qian, Z. Luo, T. Zhu, Y. Zhong, Y. Tang, M. Kong, Y. Wang, S. Jiao et al., "A survey on vision-language-action models for autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 4524–4536.

[183] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, and L. Zhou, "Neuro-symbolic approaches in artificial intelligence," National Science Review, vol. 9, no. 6, p. nwac035, 2022.

[184] J. Xu, H. Fei, L. Pan, Q. Liu, M.-L. Lee, and W. Hsu, "Faithful logical reasoning via symbolic chain-of-thought," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 13 326–13 365.

[185] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," arXiv preprint arXiv:1805.12152, 2018.

[186] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in 12th USENIX workshop on offensive technologies (WOOT 18), 2018.

[187] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model," in European Conference on Computer Vision. Springer, 2024, pp. 469–485.

[188] Y. Yang, J. Zhan, W. Jiang, Y. Jiang, and A. Yu, "Neural architecture search for resource constrained hardware devices: A survey," IET Cyber-Physical Systems: Theory & Applications, vol. 8, no. 3, pp. 149–159, 2023.
[189] R.-J. Zhu, Z. Wang, L. Gilpin, and J. K. Eshraghian, "Autonomous driving with spiking neural networks," Advances in Neural Information Processing Systems, vol. 37, pp. 136 782–136 804, 2024.
[190] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," IEEE robotics and automation letters, vol. 7, no. 4, pp. 10 914–10 921, 2022.
[191] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," Advances in neural information processing systems, vol. 35, pp. 4874–4886, 2022.
[192] C. Zhuang, H. Zhao, J. Wang, and S. Hu, "Plane constraints aided multi-vehicle cooperative positioning using factor graph optimization," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 11, pp. 17 505–17 519, 2024.
[193] B. Nebel, T. Bolander, T. Engesser, and R. Mattmüller, "Implicitly coordinated multi-agent path finding under destination uncertainty: Success guarantees and computational complexity," Journal of Artificial Intelligence Research, vol. 64, pp. 497–527, 2019.
[194] P. Sarkar, M. Hassan, A. H. Rambe, and M. Jusoh, "Integrating quantum computing into intelligent transportation systems: A comprehensive review of emerging trends, open challenges, and future research directions," Archives of Computational Methods in Engineering, pp. 1–56, 2026.
[195] S. D. Yerra, D. K. Kumar, G. K. Sheelam, R. Praveen, P. M. Paul et al., "Enhancing road safety and network intelligence using a swarm intelligence–svm hybrid model in 6g-enabled v2x communication," in 2026 IEEE International Conference on AI Engineering and Innovations (AIEI). IEEE, 2026, pp. 1–6.
[196] T. Roughgarden and É. Tardos, "How bad is selfish routing?" Journal of the ACM (JACM), vol. 49, no. 2, pp. 236–259, 2002.
[197] A. Matin and H. Dia, "Impacts of connected and automated vehicles on road safety and efficiency: A systematic literature review," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 3, pp. 2705–2736, 2022.
[198] X. Shi, L. Zhao, N. Lin, L. Xu, A. Hawbani, and Y. Bi, "Constructing future vehicle digital twins with insights from cloud-edge collaboration," in 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, 2023, pp. 110–117.
[199] M. N. Sharath and B. Mehran, "A literature review of performance metrics of automated driving systems for on-road vehicles," Frontiers in Future Transportation, vol. 2, p. 759125, 2021.
[200] W. Long, W. Wang, and K. Jin, "System-level evaluation of autonomous vehicle lane deployment strategies under mixed traffic flow," Systems, vol. 13, no. 11, p. 958, 2025.
[201] E. Fisher, M. Roberts, and O. Zelenko, "The impact of autonomous vehicles on urban mobility," Journal of Innovative Transportation Research and Development, vol. 1, no. 1, pp. 39–46, 2023.
[202] M. B. Mollah, J. Zhao, D. Niyato, Y. L. Guan, C. Yuen, S. Sun, K.-Y. Lam, and L. H. Koh, "Blockchain for the internet of vehicles towards intelligent transportation systems: A survey," IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4157–4185, 2020.
[203] C. Oham, S. S. Kanhere, R. Jurdak, and S. Jha, "A blockchain based liability attribution framework for autonomous vehicles," arXiv preprint arXiv:1802.05050, 2018.
[204] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, "Testing of autonomous driving systems: where are we and where should we go?" in Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 31–43.
[205] A. Niaz, M. U. Shoukat, Y. Jia, S. Khan, F. Niaz, and M. U. Raza, "Autonomous driving test method based on digital twin: A survey," in 2021 International conference on computing, electronic and electrical engineering (ICE Cube). IEEE, 2021, pp. 1–7.



Chengcheng Wang is currently working toward the Ph.D. degree at the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China. Her research interests mainly include autonomous driving and vehicular edge computing.



Yuansheng Song is currently pursuing a master's degree at the University of Electronic Science and Technology of China. His research interests include autonomous driving and computer vision.



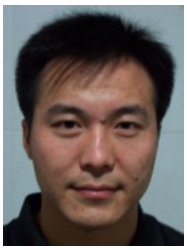
Jiahao Zhang is currently pursuing the M.S. degree in computer technology at University of Electronic Science and Technology of China, Chengdu, China. His research interests include mobile edge computing, autonomous driving and reinforcement learning.



Liang Zhao (S'09-M'17) is a Professor at Shenyang Aerospace University, China, with a Ph.D. from the School of Computing, Edinburgh Napier University (2011). With 150+ publications, his research focuses on ITS, VANET, WMN, and SDN. Liang has chaired conferences like 2022 IEEE BigDataSE and 2021 IEEE TrustCom. Serving as Associate Editor for Frontiers in Communications and Networking, and Journal of Circuits, Systems, and Computers, he received Best/Outstanding Paper Awards at 2015 IEEE IUCC, 2020 IEEE ISPA, 2022 IEEE EUC, and 2013 ACM MoMM.



Chen Chen (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in telecommunication from Xidian University, Xi'an, China, in 2000, 2006, and 2008, respectively. He is currently a Professor with the Department of Telecommunication at Xidian University and a member of the State Key Laboratory of Integrated Service Networks at Xidian University. He was awarded the Leader Talent in Innovation and Research of Shannxi Province, China, in 2022. He has authored/co-authored 3 books, and over 150 scientific papers in international journals and conference proceedings. He has contributed to developing 15 copyrighted software systems and invented over 200 patents.



Xiaoheng Deng (Senior Member, IEEE) received the Ph.D. degree in computer science from Central South University, Changsha, Hunan, China, in 2005. Since 2006, he has been an Associate Professor and then a Full Professor with the Department of Communication Engineering, Central South University, where he is currently a Joint Professor with the Shenzhen Research Institute and the Director of the Data Sensing and Switching Equipment Provincial Engineering Center. His research interests include edge computing, the Internet of Things, wireless networking and communication, data mining, and pattern recognition. He is a Senior Member of CCF, a member of CCF Pervasive Computing Council, and a member of ACM. He has been the Chair of CCF YOCSEF CHANGSHA from 2009 to 2010.



Fangming Liu (Senior Member, IEEE) received the B.Eng. degree from the Tsinghua University, Beijing, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong. He is currently a Full Professor at the Huazhong University of Science and Technology, Wuhan, China. His research interests include cloud computing and edge computing, data center and green computing, SDN/NFV/5G, and applied ML/AI. He received the National Natural Science Fund (NSFC) for Excellent Young Scholars and the National Program Special Support for Top-Notch Young Professionals. He is a recipient of the Best Paper Award of IEEE/ACM IWQoS 2019, ACM e-Energy 2018 and IEEE GLOBECOM 2011, the First Class Prize of Natural Science of the Ministry of Education in China, as well as the Second Class Prize of National Natural Science Award in China



Shaohua Wan (Senior Member, IEEE) received Ph.D. degree from School of Computer, Wuhan University in 2010. He is currently a full Professor with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China. From 2016 to 2017, he was a visiting professor at the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. His main research interests include deep learning for Internet of Things. He is an author of over 150 peer-reviewed research papers and books, including over 50 IEEE/ACM Transactions papers such as TMC, JSAC, TWC, TMM, TOSN, TODAES, TECS, TCOM, etc., and many top conference papers in the fields of edge intelligence.