Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/135/2025.20958

From Vision to Precision: Enhancing Object Detection Robustness in Autonomous Driving

Hengchen Cao[1,a,*]

[1]*University of Edinburgh, Edinburgh, EH8 9YL, The United Kingdom*
a. H.Cao-6@sms.ed.ac.uk
**corresponding author*

Abstract: In the ever-evolving landscape of autonomous driving, object detection serves as the backbone of perception, dictating the safety and reliability of entire systems. Yet, navigating the complexities of real-world environments—ranging from adverse weather and occlusions to sensor noise and unknown objects—remains a formidable challenge. These obstacles underscore the urgent need to enhance the robustness of object detection systems, a cornerstone for the advancement of autonomous driving technologies. This survey delves into the latest research aimed at strengthening object detection robustness and explores critical aspects, such as advanced data augmentation methods, resilient model architectures, multi-modal feature representation, and emerging learning paradigms. Large-scale pre-trained models, comprehensive evaluation metrics, and testing protocols are also discussed to assess robustness under diverse conditions. By synthesizing existing research, this paper identifies current gaps and proposes pathways to balance performance and robustness while ensuring scalability. This work provides actionable insights for researchers and engineers, aiming to inspire the development of safer, more reliable, and adaptive object detection technologies for autonomous driving.

Keywords: Autonomous driving, Object detection, Robustness, Deep learning, Vision-language models

1. Introduction

The rapid advancement of autonomous driving technology is transforming transportation systems, with object detection playing a crucial role in perception. The performance of object detection systems directly impacts the safety and reliability of autonomous vehicles. However, these systems face challenges in real-world environments, such as adverse weather, occlusions, sensor noise, and unknown objects, which can lead to detection failures and safety risks. Enhancing the robustness of object detection has therefore become a critical focus in autonomous driving research.

Robustness in object detection for autonomous driving means the ability of detection systems to maintain performance under various challenging conditions, such as environmental changes, sensor noise, and occlusions. It enhances the autonomous driving system's ability to make accurate real-time decisions, which is vital for the safety and ability of autonomous vehicles.

This paper reviews recent progress in improving object detection robustness, covering key areas such as data augmentation, robust model architectures, feature representation techniques, evaluation methods, and emerging learning paradigms like few-shot and meta-learning. It also explores the

DOI: 10.54254/2755-2721/135/2025.20958

potential of large-scale pre-trained models, such as vision-language models, in addressing robustness challenges. Additionally, the survey discusses open questions, including balancing model performance and robustness and designing scalable optimization methods.

By analysing the effectiveness, limitations, and complementarities of existing approaches, this survey provides a comprehensive understanding of the current state of robustness research in object detection. It aims to inspire future advancements, promoting the development of safer, more reliable, and robust autonomous driving technologies.

2. Challenges in Object Detection Robustness in Autonomous Driving

2.1. Adverse Weather Conditions

Harsh weather conditions significantly impact perception systems, impairing the effectiveness of sensors and algorithms responsible for detecting objects and interpreting the environment. For example, reduced visibility in foggy or rainy conditions can result in considerable errors in identifying obstacles or lane markings [1].

2.2. Complex and Dynamic Scene Processing

Sophisticated processing capabilities are required to accurately interpretscales or resolution scenes with multiple objects and changing environmental patterns, such as moving pedestrians or vehicles [2]. Overlapping of objects can lead to misinterpretation due to occlusion or proximity challenges. Dynamic scenes, such as moving pedestrians or vehicles, can overwhelm perception algorithms easily as they must process real-time data from various sensors [3]. Limitations in sensors also affect the performance in complex scenes [4].

2.3. Adaptation to Rare Situations

Perception systems must also handle rare events not covered during training, often referred to as the "long-tail problem".

3. Methods to Improve Object Detection Robustness in Autonomous Driving

3.1. Data Augmentation Methods

The data augmentation method artificially creates modified versions of existing data to increase the size and diversity of a training dataset. It helps improve robustness in several different ways. Here are some examples of data augmentation.

3.1.1. Generative Adversarial Networks (GANs)

GANs produce data that can simulate adverse conditions not included in the original dataset, such as images representing harsh weather situations (like rain or fog) or different periods of the day. This approach enhances dataset balance, contributing to improved model performance and increased robustness when handling diverse environmental conditions [5].

3.1.2. Semantic Domain Adaptation

Mukherjee et al. first introduced the Generative Semantic Domain Adaptation approach in their work, aiming to achieve effective semantic feature transfer and alignment between different domains. This method significantly improves perception performance in target domains such as varying driving environments or sensor setups. It leverages attribute-conditioned generative models to create

DOI: 10.54254/2755-2721/135/2025.20958

semantically varied training data, enhancing the model's generalization and performance in tasks like object classification and detection [5].

3.1.3. Adversarial Training

This method augments training data with adversarial examples. In other words, images are modified on purpose to confuse the model. It enables the perception system to identify objects despite perturbations or corruptions, enhancing robustness to unexpected input variations. An innovative adversarial differentiable data augmentation framework has been proposed by Shu et al. To improve model performance under challenging conditions, their method generates "worst-case" image transformations during training with the help of Projected Gradient Descent (PGD). This approach achieves better generalization across unseen environments than traditional techniques [6].

3.2. Model Architecture Methods

The use of different structures and designs of deep learning models is highly beneficial for improving system robustness.

3.2.1. Multi-Scale Feature Fusion

Multi-Scale Feature Fusion encompasses integrating data of different scales or resolutions from an image or a sensor to improve object detection performance. It improves accuracy in capturing objects of different sizes and helps gain a more comprehensive understanding of the scene context. An example of application in autonomous driving is MS3D, a unified LiDAR segmentation model that utilizes multi-scale feature fusion to enhance robustness and generalizability in object detection tasks relevant to autonomous driving, as discussed by Yang et al. [7].

3.2.2. Attention Mechanisms

Attention mechanisms strengthen the model’s ability to prioritize important features by focusing on specific parts of the input data. It is important in challenging driving conditions, such as low visibility or cluttered environments as it reduces false positives and improves overall detection accuracy. It also allows changing focus adaptively in response to sudden changes in the driving scene, increasing the robustness. An example of using attention mechanisms is the Multi-scale Temporal Fusion Transformer (MTFT) by Liu et al., which employs a Multi-scale Attention Head (MAH) to capture motion representations across temporal granularities and mitigate missing data issues. Combined with the CRMF module, it integrates detailed and overall motion trends, enabling robust and accurate vehicle trajectory prediction. This method achieves a 39% performance improvement on the HighD dataset, demonstrating its effectiveness [8].

3.2.3. Graph Neural Networks (GNNs)

GNNs capture relationships between entities (nodes) through edges for data represented as graphs. GNNs are highly effective at capturing and interpreting intricate relationships among objects within a scene. GNNs are suitable for processing sensor data that may not fit traditional grid formats, which improves robustness by allowing the model to effectively interpret diverse data types under varying conditions. An example of using GNNs for robust multi-modal perception is the Condition-Aware Multi-modal Fusion (CAFuser) method proposed by Brödermann et al. This approach incorporates graph-based techniques to classify environmental conditions and generate a Condition Token, enabling dynamic sensor fusion. By aligning diverse sensor inputs through modality-specific adapters and leveraging the complementary strengths of multiple sensors, CAFuser significantly improves

DOI: 10.54254/2755-2721/135/2025.20958

robustness and accuracy in adverse conditions, setting a new state-of-the-art in multi-modal semantic segmentation.

3.3. Feature Representation Methods

3.3.1. Multi-modal Feature Fusion

The integration of multi-modal data, such as visual inputs from cameras and spatial data from LiDAR, provides a robust perception of the environment. By exploiting the strengths of each sensor, this approach enhances system robustness, particularly in handling the uncertainties of complex driving environments [9].

3.3.2. Self-supervised Learning

Self-supervised learning trains models to infer missing parts of the input data from available information, encouraging a more comprehensive understanding of feature representations. This method improves robustness by reducing reliance on labelled data, equipping models to adapt to varying and challenging driving conditions [10].

3.3.3. Contrastive Learning Approaches

These methods compare similar and dissimilar data pairs, enabling the model to distinguish between objects or situations more effectively. This approach enhances robustness, particularly in scenarios with unclear or missing labels [11].

3.4. Evaluation Methods

To improve the robustness of object detection in autonomous driving, comprehensive evaluation metrics and testing protocols are crucial. This section explores methods that enhance system reliability.

3.4.1. Evaluation Metrics

Evaluation metrics play a critical role in assessing object detection performance and robustness. Intersection over Union (IoU) measures the overlap between predicted and ground truth bounding boxes, ensuring precise localization crucial for autonomous driving. Mean Average Precision (mAP) provides a single value summarizing model accuracy across classes and confidence levels, facilitating comparisons. The 3D Robustness Metric evaluates resilience under distortions through number, classification, and position robustness, ensuring reliable performance in real-world scenarios [12].

3.4.2. Testing Protocols

Testing protocols are essential for evaluating the robustness of object detection systems. Adversarial sample testing exposes models to intentionally crafted inputs designed to confuse them, identifying vulnerabilities and improving resilience against attacks [13]. Domain generalization assessment measures model performance across diverse environmental conditions without retraining, emphasizing the importance of geographically varied datasets for autonomous vehicles operating in different climates and terrains [14]. Additionally, robustness benchmarks like COCO-O test models under natural distribution shifts, such as occlusion and illumination changes, ensureand few examples for rare ones consistent performance in dynamic driving environments [15].

DOI: 10.54254/2755-2721/135/2025.20958

3.5. Large Model Applications

Large pre-trained models significantly enhance object detection systems in autonomous driving by leveraging their advanced capabilities across multiple dimensions. They improve feature learning by utilizing extensive datasets to develop rich and diverse representations, enabling better generalization across various conditions, such as different lighting, weather, and urban environments. For instance, the DriveWorld model achieved a 7.5% increase in mean Average Precision (mAP) for 3D object detection when pre-trained on a comprehensive dataset [16]. Additionally, vision-language models (VLMs) integrate visual and language data to enhance contextual comprehension, leading to improved object localization and recognition [17]. Multi-view processing architectures further bolster environmental awareness by accurately detecting objects from multiple perspectives, crucial for navigating complex driving scenarios [17]. Moreover, these models demonstrate robustness against adversarial conditions by learning from diverse datasets, ensuring safety and reliability under unexpected inputs [18]. Finally, advancements in unsupervised learning enable continuous performance improvements without requiring extensive labelled data, enhancing adaptability in dynamic environments and achieving state-of-the-art results in 3D perception tasks [19].

3.6. Learning Paradigms

3.6.1. Few-Shot Learning Paradigms

Few-shot learning enables models to recognize new object classes with minimal labeled data, addressing the challenge of rare objects in autonomous driving, such as emergency vehicles. By leveraging large datasets for common classes and a few examples for rare ones, this approach improves detection adaptability, enhancing safety and reliability in diverse scenarios without extensive retraining [20].

3.6.2. Meta-Learning

Meta-learning, or "learning to learn", enables models to quickly adapt to new tasks with minimal data. In object detection, it allows rapid adjustment to novel classes, critical for autonomous vehicles in untrained scenarios. This approach enhances robustness by maintaining accuracy when encountering unfamiliar objects or conditions in dynamic environments [21].

3.6.3. Continual-Learning

Continual learning helps models integrate new information without forgetting prior knowledge, essential for autonomous driving systems adapting to new objects and scenarios. Techniques like incremental few-shot learning, such as DualFusion, enable the detection of rare objects with minimal data while maintaining performance on common classes.

4. Conclusion

The future of autonomous driving technology lies in advancements such as open-set 3D object detection, interpretability of detection models, efficient hardware design, and integration into end-to-end systems. Open-set 3D detection enhances adaptability to unfamiliar objects. By maintaining flexible representations and continuously updating detection boundaries, the system remains robust under dynamic conditions. This capacity promotes greater safety and reliability in diverse real-world driving environments and conditions. Interpretability builds trust by allowing users to visualize and comprehend how autonomous detection models make decisions. Techniques like saliency maps highlight critical image regions or data features, illuminating the decision process. This transparency

DOI: 10.54254/2755-2721/135/2025.20958

fosters confidence among developers, regulators, passengers, and greater commercial viability, promoting safer and more acceptable autonomous systems.

Efficient hardware guarantees real-time data processing by optimizing parallel computations and minimizing latency. Additionally, integrating these hardware advancements into unified system architectures ensures seamless coordination among perception, prediction, and decision-making modules. This holistic approach enhances responsiveness, reliability, and overall performance in complex autonomous driving environments, fostering safer and more efficient vehicles. These developments will drive safer and more reliable autonomous driving systems.

This survey reviews advancements in enhancing the robustness of object detection for autonomous driving, addressing challenges like environmental variability, sensor noise, and adversarial attacks. Key findings highlight the effectiveness of data augmentation, robust model architectures, multi-modal feature representation, and emerging learning paradigms like few-shot and meta-learning. The integration of large-scale pre-trained models is emphasized as a promising direction. The study concludes that combining these approaches can significantly improve robustness, with future research needed to balance performance and scalability in end-to-end autonomous driving systems.

References

[1] *Kumar D., Muhammad N. (2023) Object Detection in Adverse Weather for Autonomous Driving through Data Merging and YOLOv8. Sensors (Basel). Oct 14;23(20):8471.*

[2] *Chai X.J., Ofen N., Jacobs L.F., Gabrieli J.D. (2010) Scene complexity: influence on perception, memory, and development in the medial temporal lobe. Front Hum Neurosci. Mar 5;4:21.*

[3] *Tata Elxsi. (2024) Overcoming challenges for AI-based perception systems in automated driving, Dec. 22, Available: https://www.tataelxsi.com/news-and-events/overcoming-challenges-for-ai-based-perception-systems-in-automated-driving.*

[4] *Matos, Francisco, Jorge Bernardino, João Durães, and João Cunha. (2024). A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions" Sensors 24, no. 16: 5108.*

[5] *A. Mukherjee, A. Joshi, A. Sharma, C. Hegde, and S. Sarkar. (2022) Generative semantic domain adaptation for perception in autonomous driving, Journal of Big Data Analytics in Transportation, vol. 4, pp. 103–117. doi: 10. 1007/s42421-022-00057-4.*

[6] *M. Shu, Y. Shen, M. C. Lin, and T. (2021) Goldstein. Adversarial Differentiable Data Augmentation for Autonomous Systems, ” in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1–8.*

[7] *Ying Li, Wupeng Zhuang, Guangsong Yang. (2024). MS3D: A Multi-Scale Feature Fusion 3D Object Detection Method for Autonomous Driving Applications. Applied Sciences, (14):10667. 10.3390/app142210667.*

[8] *Z. Liu, C. Li, Y. Wang, N. Yang, X. Fan, J. Ma, and X. Zhao. (2024) Multi-scale Temporal Fusion Transformer for Incomplete Vehicle Trajectory Prediction, arXiv preprint arXiv:2409.00904, Sep. Available: https://doi.org/10. 48550/arXiv.2409.00904*

[9] *T. Brödermann, C. Sakaridis, Y. Fu, and L. Van Gool. (2024) Condition-Aware Multimodal Fusion for Robust Semantic Perception of Driving Scenes, arXiv preprint arXiv:2410.10791, Oct. Available: https://doi.org/10.48550/ arXiv.2410.10791*

[10] *A. Prakash, K. Chitta, and A. Geiger. (2021) Multi-Modal Fusion Transformer for End-to-End Autonomous Driving, arXiv preprint arXiv:2104.09224, Apr. Available: https://doi.org/10.48550/arXiv.2104.09224.*

[11] *K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li. (2024) Multi-modal Sensor Fusion for Auto Driving Perception: A Survey, arXiv preprint arXiv:2202.02703v3, Dec. Available: https://doi.org/10.48550/arXiv.2202.02703*

[12] *X. Bi, H. Gao, H. Chen, P. Wang, and C. Ma. (2022) Evaluating the Robustness of Object Detection in Autonomous Driving Systems. The 2022 9th International Conference on Dependable Systems and Their Applications (DSA), pp. 645–649.*

[13] *M. Alimov and T. Meiramkhanov. (2024) Domain Generalization in Autonomous Driving: Evaluating YOLOv8s, RT-DETR, and YOLO-NAS with the ROAD-Almaty Dataset, arXiv preprint arXiv:2412.12349v1, Dec. Available: https://doi.org/10.48550/arXiv.2412.12349*

[14] *P. C. Chhipa, K. De, M. S. Chhipa, et al. (2024) Open-Vocabulary Object Detectors: Robustness Challenges under Distribution Shifts, arXiv preprint arXiv:2405.14874v4, Sep.*

[15] *C. Min, D. Zhao, L. Xiao, J. Zhao, et al. (2024) DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous Driving, The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15522–15532.*

DOI: 10.54254/2755-2721/135/2025.20958

[16] *L. He, Y. Sun, S. Wu, J. Liu, and X. Huang. (2024) Integrating Object Detection Modality into Visual Language Models for Enhanced Autonomous Driving Agents, arXiv preprint arXiv:2411.05898v1, Nov. Available: https://doi.org/10.48550/arXiv.2411.05898*

[17] *J. Han, X. Liang, H. Xu, et al. (2021) SODA10M: A Large-Scale 2D Self/Semi-Supervised Object Detection Dataset for Autonomous Driving, arXiv preprint arXiv:2106.11118v3, Nov.*

[18] *M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, and D. Anguelov. (2023) Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 8602–8612.*

[19] *J. Liu, X. Dong, S. Zhao, and J. Shen. (2023) Generalized Few-Shot 3D Object Detection of LiDAR Point Cloud for Autonomous Driving, arXiv preprint arXiv:2302.03914v1, Feb.*

[20] *T. Sinha. (2023) Zero-shot and Few-shot Learning for 3D Object Detection Using Language Models, MSc Thesis, Delft University of Technology, Delft, Netherlands, Nov. Available: Zero-shot and few-shot learning for 3D Object Detection Using Language Model | TU Delft Repository.*

[21] *A. Tambwekar, K. Agrawal, A. Majee, and A. Subramanian. (2021) Few-Shot Batch Incremental Road Object Detection via Detector Fusion, in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3070–3077.*