# Data-Driven Autonomous Driving Simulation: Survey

Jayesh Choudhary and Arhan Vora

***Abstract*—The Data Driven Autonomous Driving Simulation plays a significant role in development of Autonomous driving as real world testing of autonomous vehicles is very expensive as well as dangerous which presents technical and safety threats along with regulatory challenges.The data driven autonomous driving simulation helps greatly in reducing the expense and safely test and validate the vehicles and thus refine the autonomous vehicles for better and faster growth fulfilling its purpose.**

**A major problem in advancement of autonomous driving is the upper bound of autonomous driving algorithm's performance, especially in handling extreme and complex driving scenarios due to lack of real world data for such conditions as well as the lack of detailing in the data for such critical cases since the data is generally collected only from cameras instead of using the data from various sensors. The key to overcome this upper bound lies in the data-centric autonomous driving technologies which emphasises the importance of high quality data for diverse conditions in optimizing current algorithms and collecting and utilizing autonomous driving data along with dynamically upgrading it. The recent advancement in AD simulation, closed loop model training and the increase in data has been a valuable experience. However, the lack of systematic knowledge and deep understanding regarding how to built efficient data-centric autonomous driving technology are still major research gaps.**

**In this project we understand the state of data-driven autonomous driving simulation and how core technologies like artificial intelligence, machine learning, deep neural networks, sensor data fusion, synthetic data generation for rare but critical cases, integration of real-time data processing capabilities through 5G and edge computing contribute to the growth of autonomous vehicles' performance.**

***Index Terms*—Data-Driven Autonomous Driving Simulation, Real-World Testing Challenges, Sensor Data Fusion, Synthetic Data Generation, Closed-Loop Model Training, Autonomous Driving Data Collection, Performance Optimization for Self-Driving Cars, Autonomous Driving Algorithms, Simulation-Based Validation, Autonomous Driving Safety**

## I. INTRODUCTION

AUTONOMOUS vehicles (AVs) have become a major focus in research, thanks to its potential of revolutionizing transportation and enhancing road safety [1].One of the most transformative and rapidly growing technologies of the 21st century in the transportation sector is represented by autonomous driving [2]. However, real world on road testing for autonomous vehicles still remains a challenge due to its high costs [3] , safety issues, scalability, and coverage of rare edge case scenarios. In addition, regulatory hurdles add to the complexity of real-world testing. Moreover, regulatory hurdles add another layer of complexity to real-world experimentation. Data-driven simulation has hence emerged as an indispensible tool to progress the development of autonomous vehicle technology in a safer, more scalable, and cost-effective fashion to develop and validate the system.

Autonomous driving systems are extremely complex by bringing a massive amount of technological surroundings that include sensing, localization, perception, and decision making [5]. In addition, the application relies absolutely on the seamless interaction, which needs to happen between cloud platforms, huge data storages, and HD maps [5].

Though promising, autodriving technology faces a major performance hurdle given the current limitations of algorithms and the complexity of real-world driving conditions [4]. Data-driven simulation relies on the vast amount of real-world driving data from cameras, sensor inputs, artificial intelligence, machine learning models to create highly realistic virtual environments. These simulations allow developers to test autonomous driving in a vast variety of scenarios- like heavy traffic, rare conditions like natural disasters, extreme weather conditions, interaction with erratic drivers- which are difficult to test in real world. By improving algorithms in a controlled virtual environment, developers can enhance the performance of autonomous vehicle systems while reducing the danger and expense.

One of the major obstacles in the growth of AVs is the limitation of real world data, especially for rare and extreme driving situations. 90% of the autonomous driving data received is for normal driving scenarios [6]. Autonomous vehicles lack sufficient data to train their models for critical cases that do not occur often. These unusual, however critical scenarios help in training AV models to deal with the unexpected situation so that self-driving systems are secure and more reliable. In the event of an encounter with too few exposures towards these edge cases, AVs may fail to properly react at critical times. This issue is further complicated by the fact that most of training data is from lesser variety of sensors, particularly cameras leaving out essential information from other sensors [6]. This lack of diverse sensor data leaves gaps in spatial understanding and reduces the robustness of AV systems when complex driving conditions arise in front of them. Overcoming these limitations requires more data from extreme driving scenarios and the integration of multi-sensor data to form a basis for more resilient and adaptable autonomous systems.

Later progresses in high-fidelity test systems presently make it conceivable to prepare independent driving operators in closed circle, possibly circumventing inside and out the unmanageable issue of how to control the conveyance move that arises between training and deployment, and enabling scaling of training both safely and very cheaply. There is relatively little known about how to build benchmarking aids to train in

closed-loop settings. [7].

Our research focuses on the use of next generation technologies to improve and develop data-driven autonomous driving simulators, with a focus on advanced machine learning techniques in sensor fusion and on-generation synthetic data for some of the very rarest or most hazardous scenarios. Synthetic data, as such, is useful to create realistic simulations of challenging conditions that would be hard or impossible to replicate in real world experiments. Such simulation enables training in a well-controlled, safe environment and therefore increases the robustness and reliability of autonomous systems.

Emerging trends further enrich this landscape of simulations. As 5G and edge computing start to come into the picture for real-time processing, simulation fidelity, as well as the responsiveness of simulations, has been taken to new heights - enable a machine system to process and answer information more quickly and accurately. Human-AI collaboration is moreover of tall significance, since human-in-the-loop frameworks reveal imperative disclosures on intelligent between independent and human drivers to move forward more secure and more agreeable driving situations.

This development, however comes with huge challenges: knowledge and models developed in simulation must be successfully transferred to real applications. This will call for techniques like domain adaptation and generalization to ensure efficient bridging between virtual and real-world performances. There is also a need to bring to the fore the regulatory standards of simulation frameworks, which change over time, along with critical ethical questions such as fairness, bias reduction, and accountability within AI decision-making.

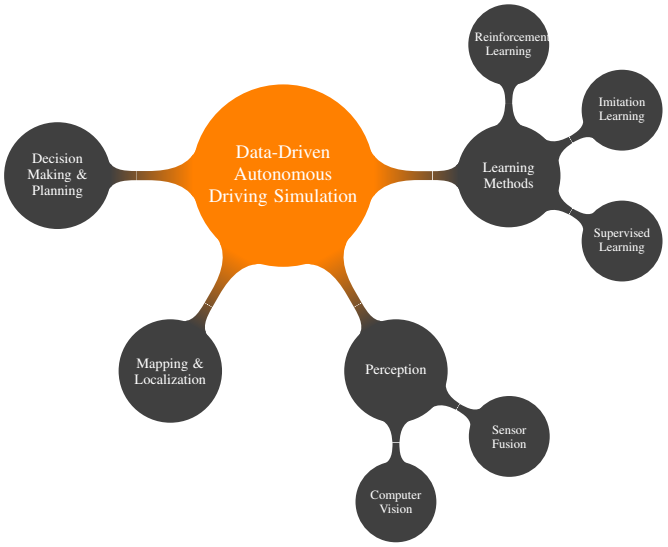


Fig. 1. Mind map of data-driven autonomous driving simulation technologies.

## II. Literature Search and Selection

### A. Search Strategy

*1) Sources and Databases:* We have used sources, including Google Scholar, IEEE Xplore, ScienceDirect, CiteSeerX, ACM Digital Library, and SpringerLink. These have many quality research papers presented in the top conference across the globe and, hence, have proven to be excellent resources for this project.

*2) GitHub:* Papers have also been cross-referenced from GitHub repositories such as Awesome Data Centric Autonomous Driving by LincanLi98, End-to-End Autonomous Driving from OpenDriveLab, and related papers from the page on SafeAV's GitHub page.

*3) Search Terms:* The keywords used to filter the search results include:

- Autonomous Driving
- Autonomous Driving Simulations
- Data-Driven Driving Simulations
- ML-based Driving
- Intelligent Driving Systems
- Algorithms for Autonomous Driving

*4) Time Frame:* Our focus was on papers published in the last decade, specifically from 2014 to 2024.

### B. Inclusion Criteria

*1) Relevance:* We are in this study to bring on board research aimed at addressing the use of different machine learning models and algorithms towards achieving autonomous driving systems. Again, the paper should share information on improving efficiency through data training on the models while the testing phase rotates around the simulations used to analyze their performance.

*2) Quality:* The articles should have huge impact and applications on real world research-projects. Refer from sources which have gone through rigorous reviewing and have been validated by experts in the field.

*3) Publication Type:* Only quality research papers and published articles that have been accepted and passed through peer review by experts in the same field are used. Also, conference papers, conclusions from reliable and trustworthy sources, and technical reports that have been verified also come within this category.

### C. Exclusion Criteria

*1) Irrelevance:* Studies which do not focus on Data Driven Autonomous Driving Simulation, it's application or reviews. The study should be aligned with the central theme of our review. Articles without a clear connection to our topic of research and those without any clear conclusion or insights should be excluded.

*2) Quality:* Exclude studies which are controversial, have proven flaws, have low credibility, showcase insufficient data. Papers which have an unclear focus, questionable assumptions, bad design, unprofessional content should be excluded.

*3) Publication Type:* Articles published by sources that are not trustworthy, non-technical publishers, articles that have not been reviewed, opinions, and incomplete research results were not be included.

### D. Screening

We started screening for study selection by searching the mentioned keywords across multiple research databases. Then

we proceeded to screen them on basis of the relevance to the topic and credibility scores. We filtered out the papers which were irrelevant to our topic and those which had low credibility. Next we filtered out those which did not prove to be up to the quality standards set by us as described above. We then started to go through these research papers and started extracting the information from these selected sources and integrated them into our own survey report providing appropriate references and citations. This approach guaranteed that we not only covered a wide range of diverse studies on the topic but also covered the topic in depth allowing us to present an in-depth comprehensive analysis report for the wide range of topics included. As a result we not only covered the breadth of the subject but also reached an appropriate depth in the subject.

## III. Simple References

## References


[1] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, *Computer vision for autonomous vehicles: Problems, datasets and state of the art*, Foundations and Trends® in Computer Graphics and Vision, vol. 12, no. 1–3, pp. 1–308, 2020.

[2] Sabiq Mirzai. The Future of Autonomous Vehicles: Revolutionizing Transportation and Society. *TechCrate*, May 24, 2023. https://techcrate.com/the-future-of-autonomous-vehicles-revolutionizing-transportation-and-society.

[3] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An Accelerated, Data-Driven Simulator for Large-Scale Autonomous Driving Research. In *Advances in Neural Information Processing Systems*, volume 36, pages 7730–7742, 2023. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/1838feeb71c4b4ea524d0df2f7074245-Paper-Datasets_and_Benchmarks.pdf.

[4] Lincan Li, Wei Shao, Wei Dong, Yijun Tian, Qiming Zhang, Kaixiang Yang, Wenjie Zhang, "Data-Centric Evolution in Autonomous Driving: A Comprehensive Survey of Big Data System, Data Mining, and Closed-Loop Technologies," Data and Knowledge Research Group, University of New South Wales, Sydney, Australia, 2023.

[5] S. Liu, J. Peng, and J.-L. Gaudiot, "Computer drive my car!", *Computer*, vol. 50, no. 1, pp. 8, 2017.

[6] Lincan Li, *Awesome Data-Centric Autonomous Driving*, Available at: https://github.com/LincanLi98/Awesome-Data-Centric-Autonomous-Driving, Accessed: September 2024.

[7] Chris Zhang, Runsheng Guo, Wenyuan Zeng, Yuwen Xiong, Binbin Dai, Rui Hu, Mengye Ren, Raquel Urtasun, *Rethinking Closed-Loop Training for Autonomous Driving*, in *Proceedings of the European Conference on Computer Vision*, Springer, 2022, pp. 264–282.