Realism Constructs for ADS Simulation Testing

Trey Woodlief
University of Virginia
Charlottesville, VA, USA
adw8dm@virginia.edu

Kevin Sullivan
University of Virginia
Charlottesville, VA, USA
sullivan@virginia.edu

Sebastian Elbaum
University of Virginia
Charlottesville, VA, USA
selbaum@virginia.edu

***Abstract*—As Automated Driving Systems (ADSs) continue to expand into the public sphere, so too must our efforts to sufficiently validate their safety. Given the wide range of scenarios over which ADSs must operate and the inherent dangers in these scenarios, developers often rely on simulation testing to exercise the system. However, the well-documented simulation-reality gap limits the transfer of results from simulation testing to real world operation, hindering the ability to build sufficient assurance cases based on validation in simulation alone. This is a fundamental issue in the construct validity of simulation-based methods for validation of ADS systems. Recent efforts have sought to decrease the simulation-reality gap through improved simulation fidelity and developing methods for generating synthetic data from real data. However, these efforts do not provide a method to assure the construct validity achieved by these improvements. Current methods to *measure* the distance between simulation and reality for ADS validation are insufficient for the task as they provide no basis on which to judge the validity of the simulated tests. For simulation testing to be trustworthy, we require methods to reason about this construct validity; i.e., whether and how much a given test or technique will yield failures that transfer to real-world deployment, or miss failures because of the lack of fidelity. We describe the continuing challenges in this domain, provide outlines of what is required of a solution, and set directions for future work in the community to this end.**

***Index Terms*—autonomous vehicles, validation, simulation testing, realism metrics**

I. Introduction

Automated Driving Systems (ADSs) promise to provide myriad benefits from increased safety, to improved efficiency, to broadened access to transportation. However, to meet these goals, ADSs must be rigorously validated over the rich space of real-world scenarios they will encounter to ensure they are sufficiently safe. Testing edge cases at the boundary of performance is vital to assure the correctness of safety-critical ADSs. Such edge cases—for example, scenarios where super-human sensing, perception, reaction, and maneuverability enable an ADS to avoid a collision that a human driver could not—are inherently unsafe to validate in the real world. To this end, simulation testing and methods that generate synthetic data for testing have become an integral part of the ADS validation pipeline due to their physical safety, controllability, speed, and cost. However, as researchers continue to develop and employ simulation and synthetic input testing, **we need methods to understand the construct validity of these tests with respect to the real world: do these tests yield insights that transfer to real-world deployment?**

We know that simulation does not match reality due to the well-documented simulation-reality gap which impacts many aspects of operating ADSs in simulation [1]–[3]. Simulation uses environment models that abstract and approximate real-world objects, leading to lower-fidelity digital twins [4]. Simulation uses models to approximate physical processes, e.g. ground friction or deformation, narrowing the range of application [1], [4]. Simulation approximates the sensor input collection process and struggles to faithfully recreate sensor artifacts and noise [5]. Each of these differences contribute to developers' concerns about the construct validity of simulation testing as a means to build a safety case for ADS real-world deployment [1], leading ADS developers to retain real-world deployment testing within their validation pipeline [6].

Yet we lack methods to *measure* such gaps between simulation and reality to reason about the construct validity of such testing efforts. Not only sensor inputs, but also actions and their effect on system and world states are approximations. Moreover, no space of digital simulations can span the space of real-world situations.

What kind of gaps in realism are relevant, not to a human but to an ADS? Does a particular test input demonstrate safety or lack thereof for the system under test? Or is executing that input a waste of resources as even if a failure is found it will not translate to deployment? Is it possible to identify this a priori to reason about construct validity? Without the ability to answer these questions, we cannot build a reliable safety assurance case through simulation testing.

To develop a robust ADS testing infrastructure that can include simulation and synthetic test data, we must answer:

1) What does it mean for an input to be *realistic*?
2) What are necessary and sufficient realism conditions for a test to substantiate an ADS safety assurance case?
3) How can we measure, compute, and decide this necessity and sufficiency?
4) How can this measurement be performed efficiently to enable practical utility in the ADS testing infrastructure?

II. Motivation

Recent research has attempted to close the simulation-reality gap for sensed inputs on two fronts: by increasing the fidelity of the simulation [3], [7]–[9], or by generating synthetic sensor inputs from inputs collected from the real world [10]–[12]. In these works, improvement in closing the simulation-reality gap is measured either through qualitative appeals to human preference on how real an input *feels* [13], or by quantitative metrics calibrated to the same human preference [14]–[18]. Many studies in this field have claimed sufficient realism, i.e., construct validity, for ADS validation by relying on these metrics, demonstrating that the average or minimum observed metric value during the study is above a given threshold. However, with a plethora of metrics that can be applied, prior techniques are evaluated over a diverse set of metrics for realism that prohibit comparison between techniques.

Further, absent a rigorous basis to connect these metrics to the safety assurance case for an ADS, i.e. to reason about construct validity, the techniques that rely on these metrics for sufficiency may prove inadequate in transitioning to field deployment: why should a score above X threshold for a particular metric be considered real enough to assure against future deployment failures of the ADS? Or it could be that a given threshold for a given metric is sufficient and *we simply do not know that*. While recent work has sought to empirically calibrate metrics to ADS performance [19], all metrics studied either did not consistently correlate with ADS performance, or required fine-tuning the metric based on the particular usage and data.

The lack of a common set of metrics for realism and a connection between these metrics and the safety case for ADS present a crucial obstacle for the field as new techniques are developed. This impedes progress toward several important research goals: test input generation techniques cannot be validated with regard to their ability to transfer to real-world deployment; test adequacy metrics cannot account for deficiencies arising from the simulation-reality gap; and simulator and synthetic test generation techniques have no sufficiency criteria to judge when they have met their realism goals and thus resources should instead be allocated elsewhere.

III. Looking Forward

We seek to call the community's attention to the need for further investigation to identify methods for measuring, computing, and deciding the necessary and sufficient realism required for the use of simulation for the ADS research and testing pipeline, if such methods exist. A useful measure to this end should enable reasoning about the sufficient realism of a test input to form the basis for a valid test. This sets several research questions for the community to address:

RQ1: *What does it mean for an input, or set of inputs, to be realistic to an ADS?*

Building from the discussion in Section II, other research fields have developed particular definitions and goals of realism based on their intended application, e.g. improving the end-user experience in virtual reality [20]. But we have no basis to infer, from human perception of realism, that an input is sufficiently realistic when presented to an ADS. The community must align on a common definition and set of goals for what it means to be sufficiently realistic as it pertains to ADS validation to determine how this aligns with the construct validity of ADS testing techniques. Only once a definition is identified can we work toward methods to measure, compute, and decide realism.

RQ2: *What parameters are required for a realism measure: the test input; the system under test (SUT); the types of faults being investigated; the method of input generation; the intended use as part of a broader assurance case?*

To begin to understand the practicality of using realism measures to decide the validity of an ADS test case, we must first understand what parameters must be considered.

For example, the application of the realism measure may depend on the sensor modality, e.g. camera versus LiDAR; the type of fault being investigated, e.g. perception misclassification versus ADS collision; or the method of input generation, e.g. simulation versus synthetic data generation among other factors. Prior empirical evidence suggests that existing simulators have demonstrated specialties, providing relatively-higher fidelity in certain aspects, e.g. improved handling of vehicle dynamics versus improved camera image generation [3], [21]; a suitable realism metric should be able to identify these differences and their applicability based on the parameterization.

An overarching realism oracle for arbitrary inputs would represent a substantial advance even beyond ADS validation, with applications in image manipulation detection [22] and deepfake detection [23]. However, as discussed, ADS test validity rests on determining whether gaps in realism are relevant *to the ADS*. Specifically targeted approaches requiring additional parameters can still provide utility. Even a measure that can be used to decide if a given input from a particular SUT targeting a particular failure-mode is realistic is useful.

RQ3: *What realism measures are suitable to this task; do different SUTs or testing paradigms require different measures or can the community build an infrastructure around one shared measure? If no shared realism measure exists, how do we demonstrate validity?*

The community has employed several measures claiming to address realism in recent years. However, without a strong, explicit argument about their connection to validity for ADS testing, no clear consensus has emerged on which measures are suitable. One of the existing measures may rise to the front, or further research may be required to develop novel measures, or identifying such a measure may be infeasible. While a common measure or set of measures would advance the community's ability to compare research, the practical utility for the end-user relies on demonstrating the construct validity of their chosen test methodology. In the absence of a suitable measure, the community must shift focus to developing alternatives for demonstrating this validity while building comparable, extensible, and valid ADS testing methodologies.

RQ4: *How can these methods be efficiently computed and pragmatically integrated into the ADS research and testing infrastructure?*

Measures that meet the prior criteria would represent a meaningful step forward in advancing research to this end. However, as an applied exercise in validating an ADS, identified methods must be amenable to efficient computation.

Prior metrics are often computationally intensive to compute, require tuning to the current task or data distribution, or rely on machine-learned components [19]. These qualities can lead to challenges such as aligning with existing software, hardware limitations, or requiring particular developer expertise. Given the complexity of the problem, this requirement must be explicitly designed for within the ADS research and testing infrastructure to permit practical use. User studies involving ADS developers can help identify core integration requirements that would enable future adoption in practice.

IV. Conclusion

In this position paper we highlight the critical need for a richer understanding of methods to reason about construct validity of simulated/synthetic input generation-based methods for ADS validation and its connection to realism and the simulation-reality gap. We outline the shortcomings of existing realism metrics that have been applied to this end and provide the contours of what this unique paradigm requires of future methods. We aim to begin the conversation in earnest for the community to build a common understanding of how we can create reliable methods for measuring realism or identify other avenues to demonstrate validity for the ADS testing pipeline. First steps in this direction may include comparing ADS behavior under similar scenarios across different simulators and real-world data to characterize the limitations of current simulators and realism measurement approaches; analyzing how ADS behaviors are affected by different input perturbations to identify ADS-relevant aspects of realism; and investigating whether, or under what conditions, existing metrics correlate with observed ADS behavior. In this way, we set the stage for future work in this direction toward building a robust infrastructure for ADS research and testing.

V. Acknowledgments

This work was supported in part by NSF Awards #2403060 and #2312487. Trey Woodlief was supported by a University of Virginia School of Engineering and Applied Science Copenhaver Charitable Trust Bicentennial Fellowship.

References

[1] A. Afzal, D. S. Katz, C. Le Goues, and C. S. Timperley, "Simulation for robotics test automation: Developer perspectives," in *2021 14th IEEE conference on software testing, verification and validation (ICST)*. IEEE, 2021, pp. 263–274.

[2] A. Stocco, B. Pulfer, and P. Tonella, "Mind the gap! a study on the transferability of virtual versus physical-world testing of autonomous driving systems," *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 1928–1940, 2022.

[3] M. Biagiola, A. Stocco, V. Riccio, and P. Tonella, "Two is better than one: digital siblings to improve autonomous driving testing," *Empirical Software Engineering*, vol. 29, no. 4, pp. 1–33, 2024.

[4] H. Choi, C. Crump, C. Duriez, A. Elmquist, G. Hager, D. Han, F. Hearl, J. Hodgins, A. Jain, F. Leve *et al.*, "On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward," *Proceedings of the National Academy of Sciences*, vol. 118, no. 1, p. e1907856118, 2021.

[5] A. Elmquist and D. Negrut, "Modeling cameras for autonomous vehicle and robot simulation: An overview," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25 547–25 560, 2021.

[6] W. LLC, "Waymo's safety methodologies and safety readiness determinations," Tech. Rep., October 2020. [Online]. Available: https://storage.googleapis.com/sdc-prod/v1/safety-report/Waymo-Safety-Methodologies-and-Readiness-Determinations.pdf

[7] Y. Li, W. Yuan, S. Zhang, W. Yan, Q. Shen, C. Wang, and M. Yang, "Choose your simulator wisely: A review on open-source simulators for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[8] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta *et al.*, "Lgsvl simulator: A high fidelity simulator for autonomous driving," in *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*. IEEE, 2020, pp. 1–6.

[9] H. Abbas, M. O'Kelly, A. Rodionova, and R. Mangharam, "Safe at any speed: A simulation-based test harness for autonomous vehicles," in *Cyber Physical Systems. Design, Modeling, and Evaluation: 7th International Workshop, CyPhy 2017, Seoul, South Korea, October 15-20, 2017, Revised Selected Papers 7*. Springer, 2019, pp. 94–106.

[10] T. Woodlief, S. Elbaum, and K. Sullivan, "Semantic image fuzzing of ai perception systems," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 1958–1969.

[11] G. Christian, T. Woodlief, and S. Elbaum, "Generating realistic and diverse tests for lidar-based perception systems," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 2604–2616.

[12] S. S. Rajan, E. Soremekun, Y. Le Traon, and S. Chattopadhyay, "Distribution-aware fairness test generation," *Journal of Systems and Software*, vol. 215, p. 112090, 2024.

[13] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

[14] Y. Wang, "Survey of objective video quality measurements," *EMC Corporation Hopkinton, MA*, vol. 1748, p. 39, 2006.

[15] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on image and Video Processing*, vol. 2014, pp. 1–32, 2014.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[17] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[19] S. C. Lambertenghi and A. Stocco, "Assessing Quality Metrics for Neural Reality Gap Input Mitigation in Autonomous Driving Testing," in *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2024, pp. 173–184. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICST60714.2024.00024

[20] G. Gonçalves, H. Coelho, P. Monteiro, M. Melo, and M. Bessa, "Systematic review of comparative studies of the impact of realism in immersive virtual experiences," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–36, 2022.

[21] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[22] S. Singh and R. Kumar, "Image forgery detection: comprehensive review of digital forensics approaches," *Journal of Computational Social Science*, pp. 1–39, 2024.

[23] Z. Akhtar, "Deepfakes generation and detection: a short survey," *Journal of Imaging*, vol. 9, no. 1, p. 18, 2023.