Received 21 October 2021; revised 3 December 2021; accepted 20 December 2021. Date of publication 5 January 2022; date of current version 24 January 2022.

Digital Object Identifier 10.1109/OJITS.2022.3140493

A Credibility Assessment Approach for Scenario-Based Virtual Testing of Automated Driving Functions

CHRISTOPH STADLER [1], FRANCESCO MONTANARI [1], WOJCIECH BARON [2], CHRISTOPH SIPPL [3], AND ANATOLI DJANATLIEV [2]

[1] Automated Driving Department, AUDI AG, 85045 Ingolstadt, Germany

[2] Computer Networks Chair of the Computer Science Department, Friedrich-Alexander University of Erlangen-Nuremberg, 91058 Erlangen, Germany

[3] Pre-Development for Automated Driving Department, AUDI AG, 85045 Ingolstadt, Germany

CORRESPONDING AUTHOR: W. BARON (e-mail: wojciech.baron@fau.de)

This work was supported in part by the AUDI AG, and in part by the German Federal Ministry of Transport and Digital Infrastructure within the Automated and Connected Driving funding program under Grant 01MM20012A (SAVeNoW).

(Christoph Stadler, Francesco Montanari, and Wojciech Baron contributed equally to this work.)

ABSTRACT An immense test space is pushing the development and testing of automated driving functions from real to virtual environments. The virtual world is provided by interconnected simulation models representing sensors, vehicle dynamics, and both static and dynamic environment. For the virtual validation of automated driving, special attention must be paid to the simulation's credibility, which can be impaired by inappropriate or inaccurate simulation models and tools. Therefore, in this work a method is proposed to assess the credibility of simulation-based testing for automated driving. The approach allows a qualitative and relatively quantitative comparisons between scenarios as well as between different simulation setups. Therefore, several uni- and multivariate metrics are applied towards a scoring of similarity of the behavior between simulation and real test drive. This is achieved by using ground truth data in form of simulation scenarios from real world measurement data. In this way, the virtual automated vehicle encounters the same conditions and surroundings than its counterpart in the real world for evaluating their similarity. The practical applicability of the proposed credibility assessment approach is demonstrated in a case study, in which the credibility of an exemplary simulation-based test bench is inferred.

INDEX TERMS Automated driving, software-in-the-loop, scenario-based approach, virtual testing, virtual development, virtual validation, computer simulation, automotive engineering, intelligent vehicles, automated vehicles.

I. INTRODUCTION

TESTING of automated driving functions (ADFs) can be performed on public roads, on proving grounds, or in a virtual environment. Testing on public roads is associated with an immense monetary effort, due to the required prototype vehicles and the training of safety drivers and co-drivers. On proving grounds, the entire environment can be controlled, but of course this resource has a limited capacity. Taking into account that several billion test kilometers have to be completed to validate ADFs [1], [2], one comes to the conclusion that this huge test space is actually impossible to manage. The idea of using the ADF to control a virtual vehicle in a simulated world emerged [3]. However, this raises the question on the credibility of the tests performed in the simulation. It cannot be denied that a gap exists in relation to real world test drives [4], [5], [6], [7], cf. Fig. 1, which is caused by the fact that simulation models can only represent the reality to a limited extent. The challenge is to quantify this gap and to answer the question whether this gap can be kept sufficiently narrow to allow credible virtual testing.

In scenario-based testing [8], [9], the virtual world is limited to a concrete setting. Other traffic participants such as

The review of this article was arranged by Associate Editor Meng Li.

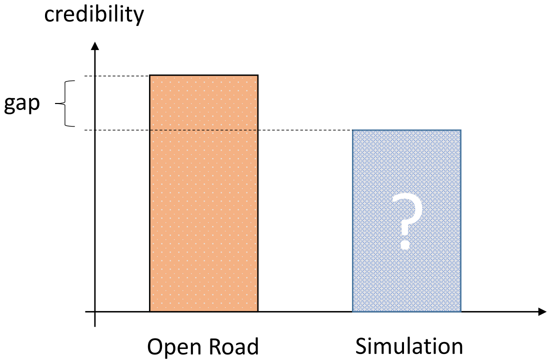


FIGURE 1. The credibility of testing approaches.

pedestrians, cyclists or car drivers, are assigned as fixed, but configurable behavior within the setting. A virtual test drive is limited to one challenge to be accomplished by the ADF, i.e., one scenario. Test scenarios can be generated in different ways [10]: manually by experts, virtually by simulations, systematically by models or extracted from real test drive data. Finally, the entire set of scenarios is collected in a scenario catalog. The basic validation idea consists of the successful completion of the entire scenario catalog by the ADF. By testing directly specific concrete traffic situations, the scenario-based approach makes it possible to be independent of the likelihood of occurrence of scenarios. As a consequence, the main motivation behind the scenario-based approach is to reduce the testing amount which results in less testing time and costs. Furthermore, by working with a scenario catalog it is possible to check which scenarios have been covered and with which rating. This eases the process for type approval. Hence, in this work the focus is on the scenario-based approach.

Nowadays, virtual scenario testing is common practice among major car manufacturers [5], [11], [12]. However, it is difficult to prove that the virtually tested scenarios are valid and can at least partly substitute testing in the real world [13]. For this reason the key question is, how to prove that simulated scenarios can be used for the validation and assessment of real automated driving vehicles. The ADF is designed and developed for functioning in the real world, but the testing is intended to be done mostly virtually in the simulation. If the same ADF behaves differently in the simulation than in the real world, the results in the simulation have no significance. Hence, it is necessary to show that an ADF behaves in a similar way to its counterpart in the real world when the same software versions are used and confronted with the same conditions in its surrounding. In this way, it is possible to reproduce the results of a physically tested scenario and to show that errors and failures can be avoided.

In this work the authors propose a new method for assessing the credibility for virtual scenario-based testing of ADFs. The idea is to perform a real test drive with an ADF and to log the bus data of the vehicle. Afterwards, the scenarios are identified and extracted from the real driving data and resimulated including the static environment and all traffic participants with the ADF connected to the simulation. Finally, the behavior of the driving function in the virtual world is analyzed with respect to the one in the real world. Depending on whether the function can reproduce its real behavior the credibility of the virtual testing can be deduced. The proposed method is a necessary preliminary work to show how representative the results of the virtual scenario-based testing are. The contributions of this work are:

- A method for assessing the credibility of a ADF simulation.
- A set of adapted metrics to quantify and qualify a similar behavior between simulation and real drive.
- A normalized relative credibility assessment for the comparison of different simulation setups.

This article is organized as follows: Section II and Section III present background information and a literature review. Section IV introduces the new approach of this work subdivided into five sections: real test drive, scenario extraction, simulation environment, simulation execution and credibility assessment. Section V presents the experimental setup with its case studies and results. The discussion and interpretation of the results follows in Section VI including its limitations. Section VII concludes and presents an outlook on future topics.

II. BACKGROUND

In this section relevant background information on the method presented in Section V is given, beginning from the in depth understanding about a scenario used within the scenario-based simulation approach and the standards applied in the case study. Furthermore, details about credibility in simulation, scenario extraction as well as the necessity of determinism are provided.

A. DEFINITION OF SCENARIO

According to Ulbrich *et al.* [14] a scenario is a temporal sequence where a development from an initial to the last scene is described. It does not only contain the consecutive frames, but also the actions, events and goals executed by all the actors and the events that are happening. In [15] the authors characterize a scenario by five layers in order to simplify the structure and the categories of information in a scenario: the static environment is defined by three layers. The first layer describes the road geometry and topology, like street dimensions with its number of driving lanes or the radius of a curve. The second layer adds traffic infrastructure, like traffic signs or traffic lights an finally the third static layer adds temporary manipulations of the first two layers like construction sites. The dynamic environment is defined in the fourth layer where all objects, e.g., vehicles, pedestrians, etc., are included with their respective interactions and maneuvers. The fifth layer describes all environmental conditions, like weather or daytime, with their effects on the first four layers.

B. ASAM STANDARDS

In recent years, the Association for Standardization of Automation and Measuring Systems (ASAM) has set many standards in the field of automotive simulation [16]. As ASAM standards have a broad basis in research and industry, these standards are solely used in this work in order to ensure the applicability of the method to the broad research community.

The ASAM OpenDRIVE (ODR)[1] simulation standard is used for the description of static elements of a road network. The standard specifies the geometry of the road as well as infrastructural elements that influence its logic, such as roadmarks, traffic signs and traffic lights.

ASAM OpenSCENARIO (OSC)[2] serves as a simulation standard for describing the dynamic entities in a traffic simulation. In an OSC file the actions and states of each individual vehicle, pedestrian, further road user and entity can be defined. OSC allows a scenario to be described in two ways, (1) based on driving actions (individual maneuvers) which consist of actions, events, goals and values or (2) based on trajectories which consist of coordinates and timestamps.

C. DEFINITION OF CREDIBILITY IN SIMULATION

There are many scales that try to measure the quality in modeling and simulation. However, the main questions is always if the simulation is credible and therefore can be used in the verification and validation process to ensure whether a real test can be replaced by a virtual one or not.

Credibility in modeling and simulation results is defined as “the quality to elicit belief or trust” according to the NASA standard 7009a [17]. Liu *et al.* [18] add that credibility has always be defined for a predetermined purpose. This also includes the data used and the results obtained from simulation models. Credibility is not directly linked to the level of quality in modeling and simulation as common verification and validation processes are, but it is strongly related to it [19]. However, metrics are needed to quantitatively evaluate the credibility assessment which define a threshold for the different levels.

D. SCENARIO GENERATION

Scenarios can be generated using different approaches. A widespread practice in the automotive industry is the generation of scenarios by experts based on their existing domain knowledge [20]. However, these scenarios are generated only by existing subjective and restricted knowledge. Thus it is challenging to prove the validity and completeness of scenarios for validation of the whole system. Therefore, it is a necessity to rely on scenarios that may and have occurred on public roads when taking into consideration the credibility of the scenarios used for validation.

Automotive companies drive a huge number of testing kilometers every day with vehicles at different stages of development [21]. Such testing vehicles record the real driving data while performing their endurance and performance runs. This results into thousands of hours of existing real driving data that can be exploited and used to understand what is happening in real traffic by identifying and extracting traffic knowledge, driving scenarios and parameter distributions. The resulting scenarios from real test drives have in fact occurred on public roads and therefore are valid and traceable test cases for the validation of ADFs.

However, the data is unlabeled, i.e., it is unknown what kind of scenarios happened during the test drive and is associated with the raw data. A scenario is not directly measurable. Nonetheless, it can be extracted by an interpretation of multiple signal patterns and their context. There exist different approaches for the identification of scenarios [22], [23], which can be subdivided into different categories: the rule-based approach starting from a scenario catalog, the supervised learning approach based on ground truth reference data and the unsupervised approach based on pattern recognition in the data. In this work a simple rule-based algorithm is presented and used in the method.

E. SOFTWARE-IN-THE-LOOP

The development and testing of ADFs in virtual environments (VEs) is carried out in so-called software-in-the-loop (SIL) simulation setups. It requires the coupling of several simulation tools and models with the ADF, which itself also consists of several software components. Fig. 2 illustrates the SIL configuration for automated driving with typical components. In general, the core of the simulation is formed by a submicroscopic traffic [24] simulator, that often contains more than just pure traffic simulation (e.g., vehicle dynamics, behavior, and sensor models). The submicroscopic traffic simulator is at times not sufficient for a reasonably realistic representation of the virtual environment and the coupling of further simulation models might be required [25], [26]. The models provide features that the standalone simulator do not yet include (e.g., network simulation) or they are custom-designed models that fulfill their function particularly well (e.g., models of specific sensors).

The operation of the ADF can be described by means of the sense-model-plan-act (SMPA) [27] methodology. The world is perceived through sensors, or in this case, sensor models. In the model step an internal accumulated world model with all known dynamic and static objects and their properties is built. The planning step is responsible for the decision-making and trajectory generation. Finally, in the actuation step set values are passed on to actuators or on to a vehicle dynamics model like in this case. The various steps typically consist of several compact software modules, which are connected to each other via diverging and merging data flow.

The distributed nature of the SIL poses some threats with respect to determinism, which might affect the repeatability of the experiments that are performed. The repeatability property is particularly important, because it forms

[1] https://www.asam.net/standards/detail/opendrive/
[2] https://www.asam.net/standards/detail/openscenario/

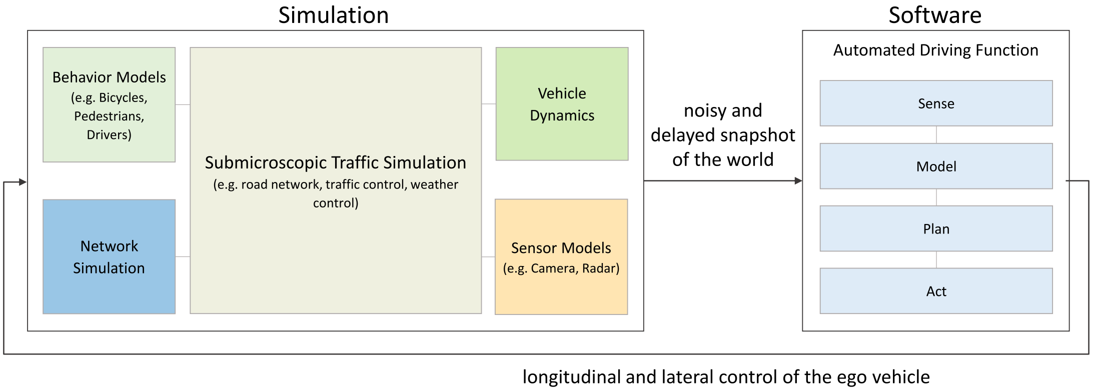

Simulation
Behavior Models (e.g. Bicycles, Pedestrians, Drivers)
Submicroscopic Traffic Simulation (e.g. road network, traffic control, weather control)
Vehicle Dynamics
Network Simulation
Sensor Models (e.g. Camera, Radar)
noisy and delayed snapshot of the world
Software
Automated Driving Function
Sense
Model
Plan
Act
longitudinal and lateral control of the ego vehicle

FIGURE 2. The SIL simulation setup for automated driving.

the foundation for regression testing and test automation. The high computational effort and the sheer number of interconnected software modules require the ADF to be executed on multiple machines and parts of it even on heterogeneous platforms, such as graphics cards. Not only it must be ensured that all software modules meet their deadlines, but attention must also be paid to data determinism. The same applies to the virtual environment. Multiple models (e.g., submicroscopic traffic simulation, sensor models, vehicle dynamics models) need to be coupled and as the simulation is computationally expensive, there might be a need to be run on multiple simulation machines. The simulation models and tools must be synchronized with respect to the simulation time. Hazards regarding non-deterministic message loss, message order and message timing need to be prohibited. In [28], [29] these phenomena are explained in more detail and synchronization mechanisms are presented, that are also used for the scope of this work. Consequently the conducted simulation runs in this work produce repeatable results.

III. RELATED WORK

In recent years different approaches for the validation and verification of ADFs have been published.

In the research project PEGASUS [30] 17 partners from scientific institutions and industry propose a scenario-based verification and validation approach for ADFs. Their central elements are based on the definition of requirements for the ADF, data processing methods, a joint and public database, the assessment of the ADF and finally the safety argumentation. This project can be seen as a first open step from a distance-based validation approach to a systematic scenario-based one. A focus of the project is the testing of scenarios on a real test track with the ability to repeat those scenarios by controlling the vehicles remotely. In this way it is possible to improve or update the ADF and analyze its behavior when confronted with the same circumstances. One realization of the project was that simulation will play a big role in the verification and validation process. Therefore, two further research projects SET Level[3] and VVM[4] arose with a stronger focus on the development of testing methods and simulation based testing.

In the research project SAVe[5] and its successor SAVeNoW[6] partners from the industry and scientific institutions develop a methodology for a holistic simulation model of a city and its traffic. The simulation model is used for the optimization of the traffic flow and the virtual testing of mobility services and ADF. The vision is to ensure a secure release of automated vehicles with the help of simulation by having a digital twin of a real digital test field in order to compare real test drive results with its virtual twins.

In [31] the authors propose different aspects to consider in order to improve validation effectiveness compared to a brute-force approach: they mention that behavioral requirements must be identified before testing the correctness in order to be able to provide pass and fail criteria. Even if the testing on the vehicle level finds no failures at all, that does not mean the ADF is necessarily safe. They state also potential solutions in order to improve the safety without the need to validate any hypothetical scenario: for instance by providing the external guarantee that the vehicle will not encounter a scenario it cannot handle. Alternatively, the ADF may contain the capability to detect that it is in a situation outside of its operational domain and put the vehicle to a safe state.

As far as official regulations are concerned, requirements for the approval of ADF are not completely defined yet. However, as a first step towards an official document in this field, there exists a proposal of the United Nations Economic Commission for Europe (UNECE) for the approval of vehicles with an automated lane keeping system (ALKS) [32]. In this document the commission presents the requirements the manufacturer must demonstrate to the technical service during the inspection for approval. It includes all the test specifications and their pass criteria, parameters and parameter ranges, operator information, data recordings during the drive, cyber security aspects and more. Simulation for the verification of the safety concept is also mentioned in

[3] https://setlevel.de/projekt

[4] https://www.vvm-projekt.de/en/

[5] https://save-in.digital/

[6] https://www.bmvi.de/SharedDocs/DE/Artikel/DG/AVF-projekte/savenow.html

particular for scenarios that are difficult to test on a test track or on public roads. However, the manufacturers have to demonstrate the scope of the specific virtually tested scenarios as well as the validity of the simulation tool chain itself.

Further framework conditions are provided by ISO26262 [33], which defines a process model together with required activities and work products as well as methods to be applied in development and production. Part 9 of ISO26262 adds special safety-oriented methods such as Automotive Safety Integrity Level (ASIL) decomposition, criteria for coexistence of elements of different ASIL classifications in a system, and requirements for safety analysis.

In addition to specific processes for the validation and verification of ADFs, there are also approaches focused on the simulation-based validation in general.

Sargent [34], Sargent and Goldsman [35] describes the challenge of validation and verification of simulation models. There are different ways to evaluate a model and different validation approaches and techniques. However, the author states that for each single problem, which should be solved with the simulation, a separate and appropriate model is needed. Therefore, there is no fixed set of methods or practices for the verification of simulation models because every model poses its own unique challenge. Some of the presented techniques, e.g., animation, comparison to other models, event validity or parameter variability-sensitivity analysis are also considered in this work.

In [36] the authors propose an approach for appropriate scenario selection and model validation for the virtual testing of ADFs. At first they present a coverage-based and data-driven approach for scenario generation for their simulation pipeline. Secondly, they show how to evaluate the behavior of the simulation model with real data with the help of statistical validation metrics and regression techniques. They state that there are deviations between simulation and reality. In addition they just use simulation as a tool, i.e., they do not evaluate or attempt to prove and improve the validity of the simulation or reduce modeling errors.

The work in [37] is based on a collected catalog of real high-severity collision scenarios. The authors reproduce those scenarios in the simulation and perform a so called "what-if" analysis by replacing one of the human-driven crash participants with an ADF. The ADF can prevent a collision in some cases. Respectively, it can at least mitigate the severity of the crash and an improvement of traffic safety can be shown.

None of the published approaches related to automated driving covers the whole simulation chain and compares the impact of the ADF in the real world with its impact in the virtual world. They are either methods on how to improve and test the ADF in the simulation or are focused on the evaluation of subsegments of the whole simulation setup. The research question in this work addresses the credibility whether real driven scenarios can be replaced by virtual driven scenarios. For this reason, the whole simulation chain is taken into account with the focus described at the end of Section I.

TABLE 1. Signal logging requirements during the real test drive.

Category	Sensor	Signal	Units
General	-	timestamp	[ms]
Ego vehicle	GPS	latitude longitude heading	[deg]
	ESC	speed	[m/s]
Infrastructure	Camera	X lane distance left Y lane distance left X lane distance right Y lane distance right	[m]
Entities	Camera/Radar/Lidar	X distance Y distance X speed Y speed class	[m] [m/s] -

IV. METHODOLOGY

When testing ADFs in simulation two important questions arise: How valid or beneficial are the tests in the virtual environment? And how can their validity or benefit be proved? This work focuses on a methodology for a valid virtual testing procedure for ADFs and indicate how to quantify and strengthen the credibility of the simulation connected to an ADF. In the following sections the single steps are presented and their importance and necessity are explained in detail. Fig. 3 summarizes the conducted steps.

A. REAL TEST DRIVE

The initial step of the proposed method is to perform a real test drive with an automated vehicle. During the driving the ADF is activated without any interruptions and is supervised by a safety driver and an additional monitoring system. The target parameters, e.g., target velocity, driving style or destination, of the function are settled. Defined data is logged during the entire trip. These measures ensure that all the required data for subsequent reproduction in the simulation is available. In addition to the target parameters of the function, the test vehicle needs to log the bus communication data in order to gather all the information from the control units and environmental sensors, e.g., cameras, radars and lidars. By this it is possible to comprehend and replicate the behavior of the test vehicle itself and of the surrounding static and dynamic entities.

Table 1 lists the minimum required signal set that must be incorporated in the logging system. Timestamp and ego vehicle signals, i.e., global positioning system (GPS) or electronic stability control (ESC), are needed for positioning the ego vehicle and understanding its driving behavior. The infrastructure signals result from the camera and are required for the detection of the lines of the driving lane. Later they are

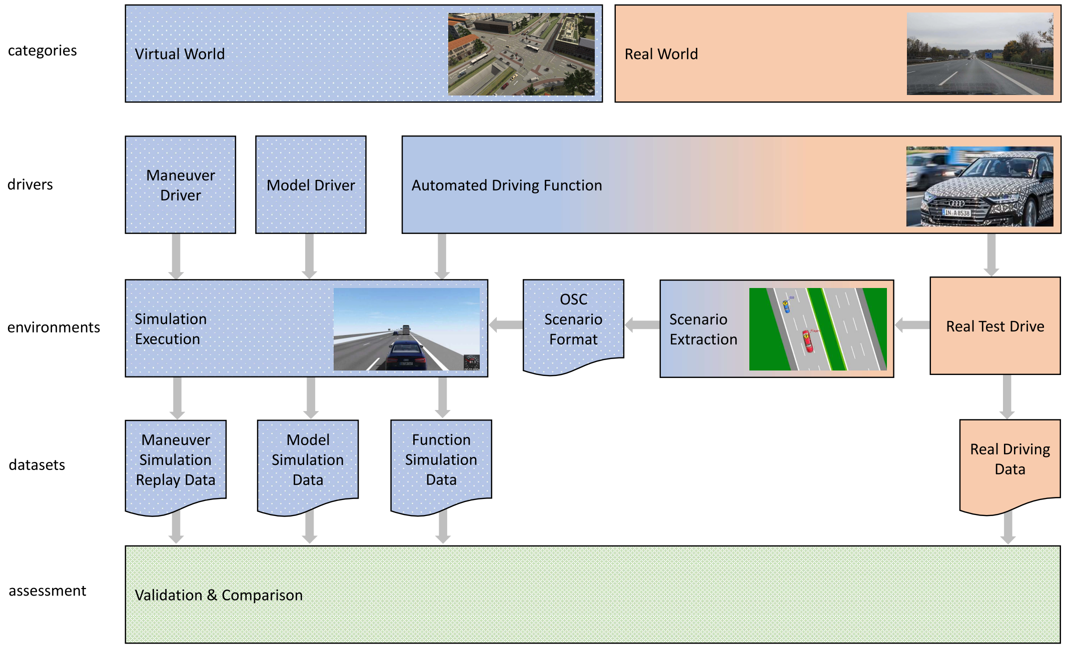


FIGURE 3. Graph of the proposed method which shows the single steps with its components.

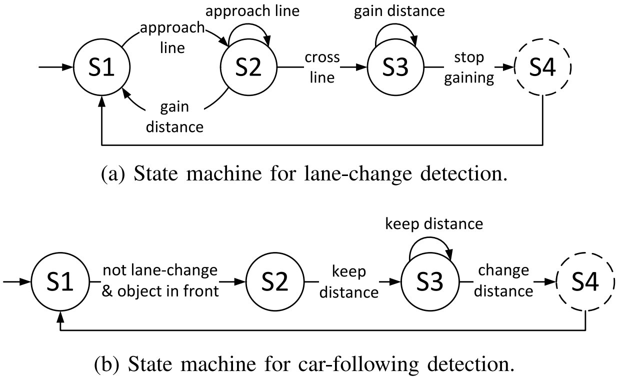
used for the identification of lane-change and cut-in maneuvers. The entity signal, which results from typical sensing sensor, are used to calculate the positioning of the surrounding vehicles relative to the ego vehicle and their respective speed and class.

B. SCENARIO EXTRACTION

The goal of this step is to take the recorded data and extract scenarios for the resimulation. As described in Section II-D the identification of scenarios from real test drive data can be performed in different ways. The aim of this work is not to establish a novel way for identifying scenarios or the creation of a comprehensive scenario catalog, but rather find sample scenarios for the validation. For the presented purpose, it is sufficient to use a rule-based method in order to find a variety of scenarios. Two different state machines help to identify lane-change and car-following scenarios on the highway. Fig. 4 shows two state machines used in this work. In Fig. 4(a) once a vehicle starts approaching a line the data is labeled as lane-change scenario between *S*2 and the ending state *S*4. In Fig. 4(b) as soon as the requirements between *S*1 and *S*2 are fulfilled once the distance is kept the labeling starts at *S*3 and ends with *S*4.

After identifying scenarios, the signal data has to be converted into simulation-tool-readable format. Thereby, the authors build on previous work that is described in [38] in detail.

At the beginning a preprocessing consisting of three steps is performed. In the first step a unique identification number (ID) is assigned to each traffic participant. Even if an ID is already assigned by the multi-sensor system, the ID can be repeated after a certain amount of time and is ambiguous. Therefore, especially for longer scenarios it is essential to clarify which detected object is related to which ID in order to avoid vehicle hopping. In the second step, an up- and downscaling of the multi-rate data from the different sensors is performed and is synchronized to a single frequency. Thereby, it is important to find a trade-off between upsampling, i.e., interpolation, of low frequency signals and downsampling, i.e., information loss, of high frequency signals. In this case the lowest common denominator of the sensor periods is applied to ensure a valid approach, because the generation of new information during the upsampling process, i.e., working with interpolated and thus not recorded data, can be seen more critical than loss in information. In the third step the algorithm detects and deletes implausible objects like, that are within the sensor's upper detection limit



(a) State machine for lane-change detection.

(b) State machine for car-following detection.

FIGURE 4. State machines used for identifying sample scenarios.

only for a short period of time, i.e., distant objects that are oscillating between being in and out of a sensor's detection range.

After having preprocessed the real test drive data, the next step is to convert the information into a simulation format. In this case it is recommended to convert the information in to the provided OSC format in order to be able to compare and exchange scenarios between different use-cases, simulation tools and development teams.

For the transformation to OSC the scenario is divided into a sequence of lateral and longitudinal maneuvers, e.g., acceleration, deceleration or lane-change. The simulation is maneuver-based, that means the virtual vehicles are assigned simulation time or position-based actions and not fixed trajectories. The advantage is, that the scenario is described in a configurable manner and can be altered in an intuitive and comprehensive way by test engineers. However, while abstracting the real test drive data into a sequence of maneuvers information is lost due to simplifications, e.g., linear acceleration.

C. SIMULATION ENVIRONMENT

Within a closed-loop simulation setup, a valid virtual environment model is a prerequisite for testing and validating ADFs, despite valid further components like sufficient sensor models and a vehicle dynamics model. The quality of the virtual environment model, also referred to as the digital twin, used for the simulation of automated driving is often not clearly specified and the generation of highly accurate and detailed models is very costly on a large scale. Nevertheless, comparability with the real world has to be ensured. In the context of driving simulation a digital twin of the surrounding environment is a virtual model with physical objects and geometries in the street space. Therefore, as a minimum requirement for driving simulation a virtual environment model of a street space is needed, which is comparable to the one in real world where the scenario actually happened. As a consequence, so called high-definition (HD) maps are available, which are based on the surveying of traffic networks with cameras and/or laser scanners in combination with precise positioning systems. These raw data are transformed in a machine readable format like ODR.

From the perspective of automated driving, the real world is a complex system that could have an infinite number of different characteristics and interactions. Additionally, simulation fidelity is not clearly defined within the generated maps. Consequently, rebuilding the digital twin by measuring the reality is almost impossible in terms of accuracy and completeness of appearing contents. For this reason, the virtual environment used for a specific application in driving simulation has to be checked whether its modeled level-of-detail is sufficient or not to assure validity. This can be achieved by identifying the relevant parameters in the virtual environment of the road network including its near surrounding. The environment model described by the ODR format is represented by a finite number of parameters. Therefore, relevant parameters can be identified, e.g., by sensitivity analysis by variation of the original ODR parameters and the investigation of the influence on the virtual drive within these samples. With further evaluations with methods like the Optimal Subpattern Assignment (OSPA) metric [39], [40], a quantification of the impact and the relevance of parameters can be achieved. This serves as a quality assessment for virtual environment models. This information can be looped back to improve accuracy if necessary of influential parameters or respectively take the known uncertainties into account for validating the whole driving simulation [41].

D. SIMULATION EXECUTION

The SIL for automated driving (AD) consists of multiple simulation units and the ADF. The key simulation unit is the submicroscopic traffic simulator. The submicroscopic traffic simulator is responsible for the representation of the virtual world including the road network. The map is loaded via the standardized ODR format. In the presented work, it is always the same file, since the test drive was limited to a bounded region. The dynamic behavior of the road users is described via the standardized OSC format. Per scenario, the higher-level simulation control takes over the loading of a new file. The dynamics of the ego vehicle are represented by an additional vehicle dynamics model. The configuration of the vehicle dynamics model was based on the characteristics of the prototype vehicle with which the real test drives were carried out. The perception of the ego vehicle is limited by several sensor models. Geometric sensor models are used for this purpose. The parameters like mounting points and fields of view are based on the prototype vehicle. The granularity of simulated perception is based on objects, not raw sensor data. There is an interface mismatch between simulation units and function units. Conversion modules handle the mapping of simulation signals into function signals and vice versa. The time progress of the simulation is not bound to real-time. The data exchange between individual modules and the distribution of simulation time to the modules is handled by a higher-level middleware that features conservative synchronization mechanisms. For the same scenario and the same parameters including the same initial values, the ego vehicle always behaves in the same way. The virtual test drives are reproducible. Therefore, a worst-case behavior analysis over several simulation runs can be omitted.

E. CREDIBILITY ASSESSMENT

The aim of the approach is to assess the credibility of the simulation results with the artefacts generated described in the steps from Section IV-A to Section IV-D so far. For this purpose the generated simulation results are compared with the real test drive data which serves as ground truth. In order to validate whether the ADF behaves equivalently in the real and virtual drive and to ensure the simulation is sufficiently detailed, a step-wise procedure is presented to ensure relative credibility of the simulated scenarios which is summarized in Fig. 5.

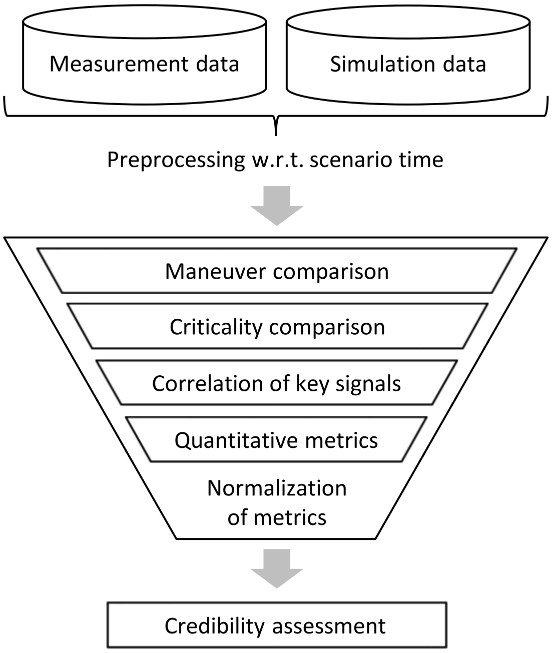


FIGURE 5. Procedure for validation of simulation data.

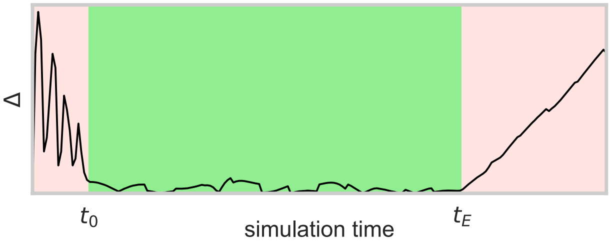


FIGURE 6. Interval for validation comparison.

As one prerequisite the quality and accuracy of the test drive data, i.e., typical noise floor originated from the installed sensors, must be known. Furthermore, the assumptions made for extraction and abstraction of the scenarios has to be taken into account as well as the completeness and fidelity of the virtual environment model. This background knowledge about the input data quality and a proper simulation execution with regards to determinism is needed to exclude potential side effects and to avoid misinterpretations in the following validation steps, when simulation data is evaluated against the data from the real test drive.

The simulation data has to be pre-processed in terms of the timestamps considered for evaluation before quantitative validation can be applied. Fig. 6 shows the graph of a plausible deviation error Δ per timestamp against ground truth data. On the one hand side, state of the art vehicle dynamics models need some time from the start of the simulation execution until they are calibrated. This is particularly the case when the simulation starts with a high initial velocity of the ego vehicle.

Therefore, simulation should only be used for validation after the end of the initialization period at t_0. On the other side, the length of the extracted scenario must be strictly satisfied. Immediately after the end of the last extracted maneuver sequence at time t_E the ADF still controls the ego vehicle, but the executed maneuvers have no reference in the test drive data and thus cannot be compared.

Before determining the absolute deviations between the real and virtual test drive, the maneuvers sequences are juxtaposed for a scenario. If the same maneuvers are driven in simulation and are triggered at approximately the same time, an equivalent behavior of the ADF can be implied.

In order to test whether the resimulation is appropriate in case of criticality, a criticality assessment of the investigated scenario has to be conducted. As a generic metric the time-to-collision (TTC) can be used to compute the time until a collision occurs, upon condition that the vehicles involved maintain their current speed according to amount and heading and that they are on a collision course. TTC at a specific time i is defined as

$$TTC_i = \frac{d_i}{v_{rel_i}} \tag{1}$$

with the distance d and the relative velocity v_{rel} between the vehicle under test and a surrounding object. In literature the TTC is defined as critical for values $TTC_{crit} \leq 1.5s$ [42]. In addition to that a multidimensional criticality analysis as presented in [43] can also be applied. A scenario can be critical (a) or not (b) in the real test drive. In case of (a) and the simulation also detects critical maneuvers, the methodology can provide a valid virtual testing regarding the criticality assessment. That is also be valid for case (b) and the resimulation being noncritical. If an opposing criticality of real and virtual behavior is observed, function testing regarding safety cannot be applied and resimulation is not credible.

In a next step, the correlation of key signals, i.e., position, velocity and acceleration, is computed to identify the signals that negatively affect the deviations calculated by the more detailed metrics in the following steps. Given two datasets x and y as paired $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ the Pearson correlation r [44] is defined as

$$r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}. \tag{2}$$

The correlation describes a linear dependence of the examined signals and is at its maximum ($r = 1$) when there is a perfectly positive relationship. Conversely, it is at a minimum ($r = -1$) when there is an anti-correlation, that is, when there is a perfectly negative relationship. $r = 0$ means that there is no linear relationship between the datasets. This helps to improve potential uncertainties of the input models as well as only translational displacements of the driven maneuvers.

In order to compare the credibility of simulation results to real test drive data, quantitative metrics are needed. There is a series of standard metrics for computer simulation validation [45]. The most widely applied metric is root mean squared error (RMSE) [44]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \|x_i - y_i\|^2}{n}}, \tag{3}$$

which computes the average deviations in form of the Euclidean distances. As $RMSE > 0$ by definition, lower values indicate a better fit between two datasets than higher ones. In the evaluations the RMSE is not used to directly compare the trajectories driven by the vehicles, instead a deeper analysis is conducted within the simulations variants to distinguish between the different sources of errors occurring.

A large number of signals are available for evaluation in each time step of the simulation. Therefore, a unique metric that in the best case represents as much information as possible in one index is desirable in order to achieve a meaningful expression of the simulation quality. Further requirements are the general applicability for all kinds of driving scenarios, the ability to combine static environment information with the vehicle movement as in a closed-loop simulation that information is directly linked to the ego vehicle's response. These requirements can be fulfilled with the OSPA metric [39], [40]. The OSPA metric consists of multiple components.

$$OSPA = \left(d_{lo}^{p} + d_{ca}^{p} + d_{la}^{p}\right)^{\frac{1}{p}}. \tag{4}$$

The first component d_{lo}^{p} contains the localization error, i.e., the distance function between two assigned objects. The second component d_{ca}^{p} corresponds to the cardinality error, i.e., the unassigned objects compared to ground truth. The labeling error component d_{la}^{p} penalizes incorrect assignments. In all components p refers to the order of the used Wasserstein metric for calculation. A more detailed computation of the three components of the OSPA metric can be found in literature [39], [40]. Accordingly, the OSPA allows the states of multiple objects to be evaluated over time.

The calculated error values provide a rating about the absolute difference between the real and virtual test drive and gives a first estimation on the quality of the simulation results in terms of credibility. Additionally, a solely maneuver simulation is conducted with an ideal model components as well as with a reference model to be able to classify the ADF test drive.

So fare the calculated values for the metrics are based on absolute deviations for specific scenarios. Therefore, a normalization of the metrics presented above is introduced in order to compare scenario independently and to allow a scoring against a threshold. For each metric its normalized value is defined as the probability $P(z_t \leq Z_{th})$ with z_t as the value of a metric at timestamp t and a threshold value Z_{th} of the corresponding metric. As thresholds are dependent on the requirements defined for a specific test case an overall credibility assessment can only be derived relatively to the thresholds. For a final decision binding for approval a link to the requirements which have to be fulfilled has to be established.

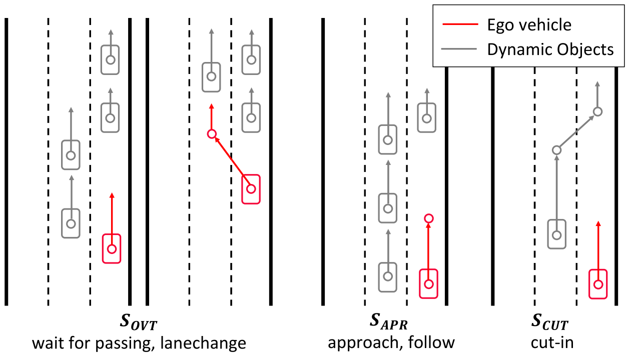


FIGURE 7. Maneuver definition for the scenarios of the conducted case study.

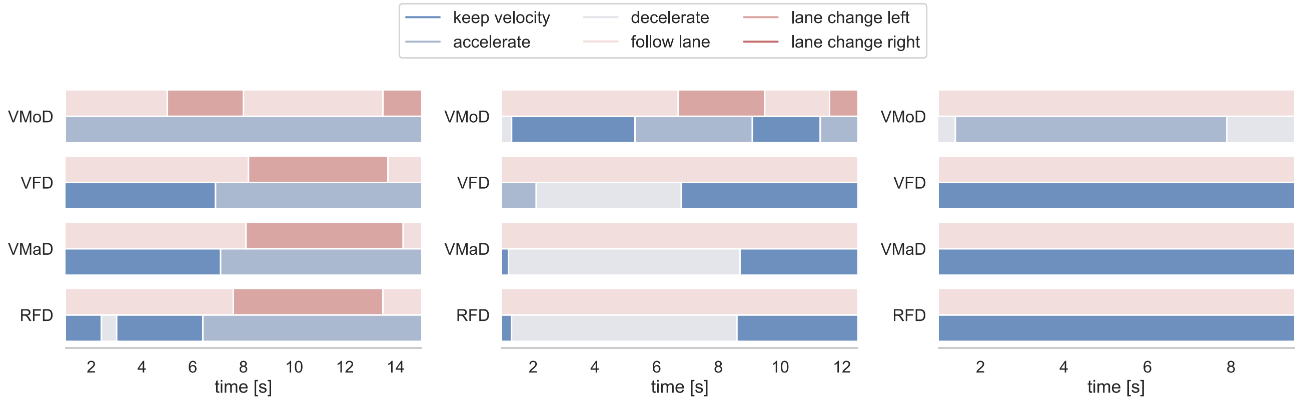
V. RESULTS

The test drive for the experiment of this work is carried out in a pre-development prototype vehicle equipped with a data logging system and an ADF. The test drive is recorded on the digital test field on the highway A9 with three lanes in the proximity of the city of Ingolstadt. For this region a pool of map data with road networks in the ODR format exists. As submicroscopic simulation tool Virtual Test Drive (VTD) [46] is used. However, the proposed methodology is based on the ASAM simulation standards and is thus compatible with any supporting simulation tool.

A large number of scenarios from the real test drive can be generated. From the variety of scenarios, three representational scenarios are selected with distinct characteristics for further investigation. With S_{OVT} a lane-change scenario and with S_{APR} a car-following scenario are chosen as representatives for lateral, respectively longitudinal control actions. Additionally a cut-in scenario is analyzed, as the reaction of the ego vehicle is solely triggered by the surrounding object and is considered as safety-critical. Fig. 7 illustrates the actions of the selected scenarios described in detail in the following:

In S_{OVT} a lateral maneuver of the ego vehicle occurs. In the beginning the ego vehicle is following a bus, but has the intention to overtake because the velocity is too low compared to the target velocity. As there is right-hand traffic in Germany, in order to overtake a lane-change to the left need to be performed. However, the lane on its left side is occupied. Therefore, the ego vehicle waits until the occupying vehicles pass by in the middle lane. As soon as the middle lane is unoccupied, the ADF releases a left lane-change and acceleration to the target velocity is triggered.

S_{APR} is a car-following scenario where the ego vehicle approaches the car in front and follows it. In this scenario there are only longitudinal maneuvers. During the whole scenario the middle lane on its left side is occupied by various vehicles. Even if there is an intention to overtake, it is impossible to change the lane because the gaps are too



(a) S_{OVT} (b) S_{APR} (c) S_{CUT}

FIGURE 8. **Comparison of the identified lateral and longitudinal maneuvers.**

short. Therefore, the ego vehicle remains on the right side and follows the vehicle in front.

In the beginning of S_{CUT} the ego vehicle is driving on a free lane with its target velocity and without any object in front. A vehicle overtakes the ego vehicle from the left side, performs a cut-in maneuver and merges into the lane of the ego vehicle. Since the safety distances are respected the ADF keeps its velocity without any hesitation. Thus, in this scenario the ego vehicle is performing only one keep velocity maneuver during the whole time and should not be affected by the happening in the surroundings.

After performing the test drive and executing the proposed steps of Section IV, four different setups to evaluate the success and the credibility of the test scenarios in the simulation are introduced and compared:

- The real test drive of the prototype automated vehicle driving in the real world, which refers to as Real Function Drive (RFD).
- The virtual test drive data of the virtual automated vehicle driving in the virtual world, which refers to as Virtual Function Drive (VFD).
- The virtual maneuver-based OSC drive, where the ego vehicle retraces extracted maneuvers, which refers to as Virtual Maneuver Drive (VMaD).
- The virtual model-based drive, where the ego vehicle is controlled by a tool-specific driver model, which refers to as Virtual Model Drive (VMoD).

In the next sections the evaluation is divided into six steps categorized from abstract to detailed. It starts from a comparison of the occurred maneuvers, secondly a criticality analysis, thirdly a correlation analysis, followed by a detailed RMSE and OSPA comparisons and ends with a normalization step.

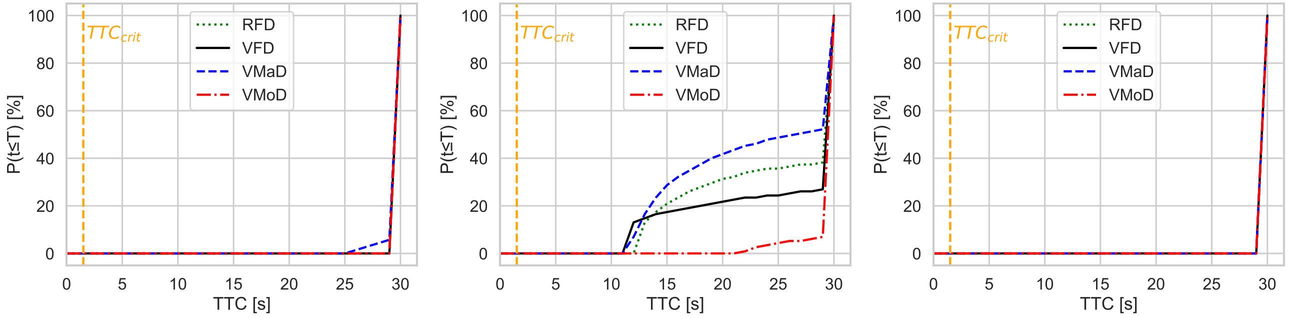
A. MANEUVER COMPARISON

First, the maneuvers performed for the four setups are determined and compared. The analysis is divided into lateral (lane-changes) and longitudinal (acceleration processes) maneuvers. The goal is that the maneuvers performed in real driving by ego vehicle are also reproduced in the simulation. This builds the basis for a similar impact of the ADF in the real and in the virtual world. A more in-depth analysis is considered to be meaningful only if there is strong correspondence concerning this matter.

For the determination of the longitudinal maneuvers acceleration values of the ego vehicle are used and for the lateral maneuvers the position data of the ego vehicle in relation to the position data of the lanes is applied. The results of this analysis are depicted in Fig. 8. In all scenarios, there is a strong correspondence between VMaD, VFD, and RFD for both longitudinal and lateral maneuvers. As expected, the simulated maneuvers also show the strongest correspondence. In S_{APR}, there is only a slight shift in the deceleration process and in S_{OVT}, there is a slight shift in the initiation of the lane change process. The lane-change also takes slightly longer. S_{CUT} completely matches for the VMaD and the VFD. The VFD shows a comparably good quality for S_{OVT}. Only the beginning and the length of the lane-change is slightly different. In S_{APR}, the VFD accelerates initially and the deceleration process is shorter. A closer look at the data reveals that the acceleration and deceleration values in this scenario are very low in terms of magnitude. In addition, the ego vehicle is very close to the threshold of the safety distance to the vehicle in front. As a consequence, a minor difference in the mapping of the position of the front vehicle already has an effect on the maneuvers performed. In the RFD a short deceleration occurs in S_{OVT}, which was not performed by any of the other variants. This is caused by a measurement error. In contrast, the VMoD that is used as reference performs significantly deviating maneuvers. In S_{OVT} and S_{APR}, for example, additional lane changes are executed that did not occur in the RFD. There is no lane change in S_{CUT}, but heavy acceleration and deceleration can be detected.

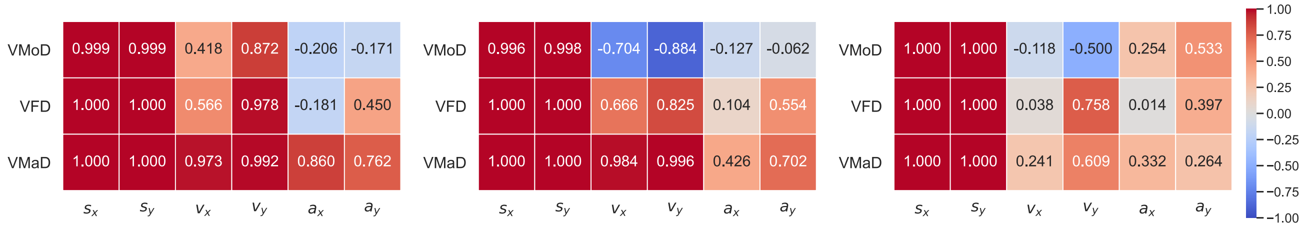
B. TTC

The second step of the credibility assessment is based on criticality in order to verify whether safety-relevant criteria



(a) S_{OVT} (b) S_{APR} (c) S_{CUT}

FIGURE 9. Criticality assessment of the scenarios.



(a) S_{OVT} (b) S_{APR} (c) S_{CUT}

FIGURE 10. The Pearson correlation coefficients of the variants.

are met in a similar way for the real and virtual drives. The TTC is applied because it is presumably the best-known and understood safety-related parameter. The TTC is determined from the perspective of the ego vehicle in every simulation time step. The evaluation is capped to a maximum TTC of $30s$. Larger values are not considered to be relevant, as they are very far from the critical threshold of $TTC_{crit} = 1.5s$. No TTC can be determined for time points without collision course. For these time points the TTC is set to the maximum value of $30s$. Thus, the ratio of critical to non-critical time points is not distorted. A cumulative distribution function (CDF) is composed from the TTC time series. The CDFs for all scenarios are illustrated in Fig. 9. For none of the three scenarios is any drive even close to the criticality threshold. In addition, there are barely any time points with collision course for scenarios S_{OVT} and S_{CUT}. In the scenario S_{APR} the ego vehicle approaches a leading vehicle, but cannot overtake due to an occupied lane. In all cases, however, an anticipated speed adjustment of the ego vehicle takes place, so that the TTC values are also in uncritical ranges above $10s$. In this process, VFD and VMaD replicate the criticality of RFD better than VMoD. Consequently, as all of the scenarios and drives show equivalent criticality and no violations with regard to TTC_{crit}, further analysis are conducted to quantify the credibility.

C. CORRELATION ANALYSIS

In the next step, the Pearson correlation coefficients are examined for the position s_i, velocity v_i, and acceleration a_i data of the ego vehicle for $i \in [x, y]$, where x denotes the longitudinal and y the lateral direction. The correlation is performed in each case in reference to the RFD. A maximum positive correlation is desirable ($r = 1$), because then the ego vehicle signals in simulation behave qualitatively identical to the ones in the RFD. The trajectories or the acceleration profile, for instance, are then similar.

At first, a check is done at which point the maximum of the cross-correlation of the individual signals is located in order to identify potential temporal shifts. The result is that there is no time lag for any of the variants or signals. The Pearson correlation factors are thus determined between the respective unshifted signals, as illustrated in Fig. 10.

The same pattern emerges for all three scenarios. Positions are in each case, for all variants, better reflected than velocities, and velocities are better reflected than accelerations. The superior results in the position data are certainly linked to the map and road layout in the simulation. Since the virtual ego vehicle does not drive off the road, the correlation values are already very high. From the velocity correlation it can be seen that in each case a similar goal is pursued by the virtual ego vehicle: namely, comfortably reaching the desired speed without increasing criticality. However, how this goal is reached is reflected in the acceleration correlations. The vehicle dynamics model also has a major influence on this. The decreasing velocity correlations could be an indication that the vehicle dynamics model could possibly be parameterized more effectively to match the dynamics of the real prototype vehicle. A comparison of the variants also shows a similar trend for all three scenarios. The VMaD shows the highest correlation, closely followed by the VFD. The VMoD performs worst for all three scenarios. Partially, there is no or even negative correlation in the velocity profiles.

TABLE 2. Results of the RMSE evaluation.

		VMaD		VFD		VMoD	
		Absolute	Relative	Absolute	Relative	Absolute	Relative
S_{OVT}	$\Delta s\ [m]$	0.99 ± 0.24	2.53 ± 0.74	1.44 ± 0.40	2.67 ± 1.23	9.28 ± 7.52	9.33 ± 5.58
	$\Delta v\ [\frac{m}{s}]$	0.23 ± 0.11	0.72 ± 0.42	0.54 ± 0.32	0.96 ± 0.35	2.77 ± 1.25	2.74 ± 1.16
	$\Delta a\ [\frac{m}{s^2}]$	0.20 ± 0.08	0.60 ± 0.51	0.59 ± 0.34	0.73 ± 0.54	0.83 ± 0.80	1.09 ± 0.78
S_{APR}	$\Delta s\ [m]$	0.91 ± 0.23	3.01 ± 1.79	3.52 ± 1.96	3.33 ± 1.88	11.37 ± 4.95	10.70 ± 3.95
	$\Delta v\ [\frac{m}{s}]$	0.24 ± 0.13	0.97 ± 0.52	1.07 ± 0.54	1.29 ± 0.44	3.37 ± 1.47	3.54 ± 1.77
	$\Delta a\ [\frac{m}{s^2}]$	0.22 ± 0.22	0.87 ± 0.77	0.94 ± 0.52	1.35 ± 0.78	1.31 ± 0.81	1.76 ± 1.01
S_{CUT}	$\Delta s\ [m]$	1.66 ± 0.84	1.17 ± 0.46	1.38 ± 0.88	0.60 ± 0.33	11.07 ± 4.23	10.51 ± 3.46
	$\Delta v\ [\frac{m}{s}]$	0.39 ± 0.06	0.59 ± 0.53	0.89 ± 0.28	0.60 ± 0.42	1.75 ± 1.04	1.39 ± 1.48
	$\Delta a\ [\frac{m}{s^2}]$	0.09 ± 0.04	0.58 ± 0.69	1.05 ± 0.50	1.15 ± 0.72	0.46 ± 0.24	0.88 ± 0.85

D. RMSE

For the first time, a look at absolute error values expressed by the RMSE is taken by calculating the RMSE for all simulation variants, all simulation scenarios and for each simulation time step. In addition, it is distinguished the error of position, velocity and acceleration (1) of the ego vehicle itself and (2) of all relative vectors to all other traffic participants located in the simulation. The relative error vectors are normalized to error per one observable traffic participant in the corresponding simulation time step.

The average RMSE and the standard deviation over all simulation time steps are illustrated in Table 2. The general trend of lowest errors for the VMaD and the highest errors for the VMoD remains. The VFD is positioned in between. As it has already been shown, deviating maneuvers are performed in the VMoD. This is reflected in a positioning error of the magnitude of $10m$. The VMaD and the VFD exhibited very similar maneuvers, albeit somewhat temporally delayed. This translated into a low single-digit error in meters. No statement can be made as to whether the absolute positioning of the ego vehicle or the relative positioning to other traffic participants is better reproduced. The same applies to velocity and acceleration errors. The high acceleration errors in the VFD are remarkable. The cause lies in the oscillation of the acceleration of the ego vehicle. All scenarios start with a high initial speed of the ego vehicle on the highway. This poses a challenge for the ADF and vehicle dynamics model, since internal states are not initialized.

E. OSPA

In order to analyze the simulation data of the trajectories over time compared to a ground truth, in our case the measurement data from the real test drive, the OSPA metric described in Section IV-E is used. On the one hand, the overall deviations occurred during a scenario ($\sum \Delta_{step}$) are evaluated in Fig. 11, and on the other hand especially for maneuver-based simulations the error entries per timestamp (Δ_{step}) allow conclusions regarding the maneuvers driven over time.

For the results of the accumulated deviations computed with the OSPA metric similar trends can be observed qualitatively for each kind of simulation data independent of the scenario. The VMaD performs best, while the VFD approximately doubles this error. A closer look at the error entry per time step shows more or less constant deviations over time but with sawtooth-shaped behavior. The sawtooth curve originates from the lower sampling rate of the GPS signal used in the real test drive compared to the high frequency driving simulation setup by a factor of ten. In contrast to quite similar and low deviations of VFD and VMaD, VMoD performs worst among the three simulation results. As a standard driving model is used for VMoD higher deviations are reasonable. High deviations can especially be observed in S_{OVT} from second seven following when a lane-change is conducted according to the maneuver definitions or in S_{CUT} when a cut-in occurs in front of the ego vehicle until second five. The described behavior can be qualitatively observed for each scenario of the case study.

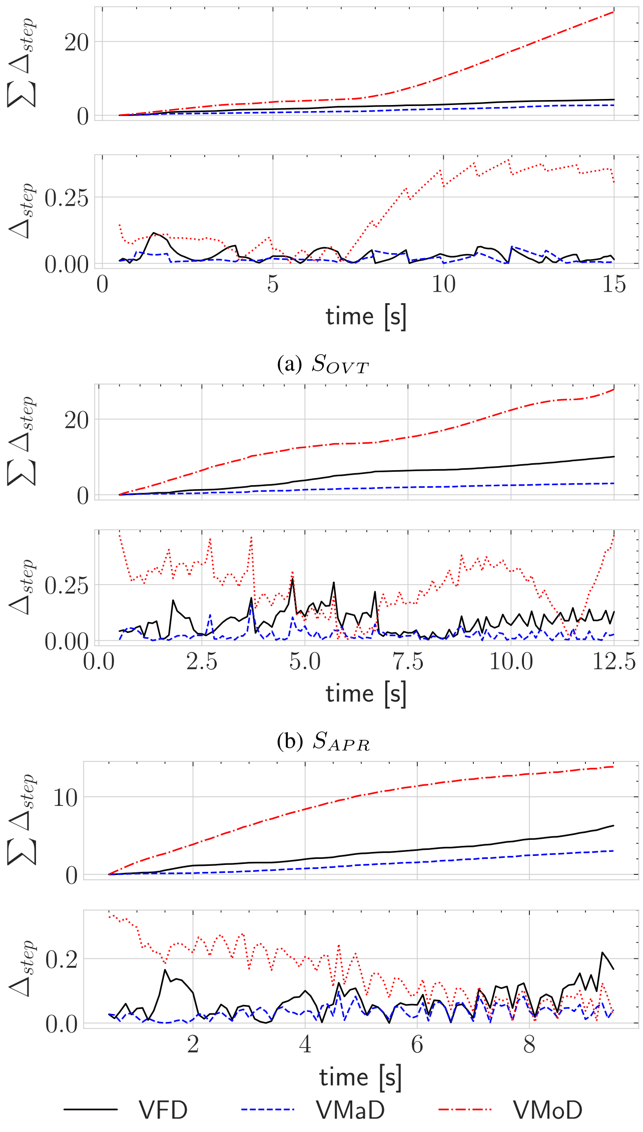
TABLE 3. Comparison of the mean and standard deviation of Δ_{step} of the OSPA metric for the case study.

	VMaD		VFD		VMoD	
	Mean	SD	Mean	SD	Mean	SD
S_{OVT}	0.019	0.014	0.030	0.023	0.20	0.14
S_{APR}	0.023	0.024	0.085	0.051	0.23	0.11
S_{CUT}	0.034	0.022	0.071	0.046	0.15	0.08

In order to evaluate the three scenarios irrespective of their duration and resimulated maneuver types, Table 3 compares the mean and standard deviation (SD) of Δ_{step} shown in Fig. 11 for the three different variants and among each scenario of the case study. The comparison is calculated per timestamp. In general, the mean as well as the SD reveal values in the order of magnitude of 10^{-2} for the VFD and VMaD. In contrast the VMoD show error deviations in the order of magnitude of 10^{-1}. Consequently, the conducted approach delivers similar results for each scenario of the case study, what emphasizes the general applicability of the analyses.

F. NORMALIZATION

In the last step the data is normalized to a range of values between 0 and 1. The value 0 corresponds to a bad credibility



(a) S_{OVT}

(b) S_{APR}

(c) S_{CUT}

FIGURE 11. Evaluation of OSPA metric per timestamp and accumulated over simulation time.

and the value 1 to the highest credibility. During normalization, particular attention is paid to the comparability of the resulting metrics both between scenarios and between entire simulation setups. Thus, it should be comprehensible for which scenarios (e.g., urban vs. rural) the simulation setup still has weaknesses. In addition, an improvement in individual models should also be reflected in a higher overall score. An evaluation of different simulation tools can also be conducted. The normalization of the individual metrics is carried out as follows.

- *Maneuvers:* the time fraction, where the executed maneuver matches the reference scenario
- *TTC:* the time fraction, where the criticality matches the reference scenario
- *Correlation:* linear mapping of the interval $[-1, 1]$ to $[0, 1]$
- *RMSE:* the time fraction, at which the physical parameters do not exceed defined threshold values
- *OSPA:* the time fraction, at which the OSPA metric does not exceed a defined threshold value

TABLE 4. Normalized OSPA metric.

	VMaD	VFD	VMoD
S_{OVT}	0.932	0.747	0.123
S_{APR}	0.884	0.322	0.075
S_{CUT}	0.967	0.550	0.132

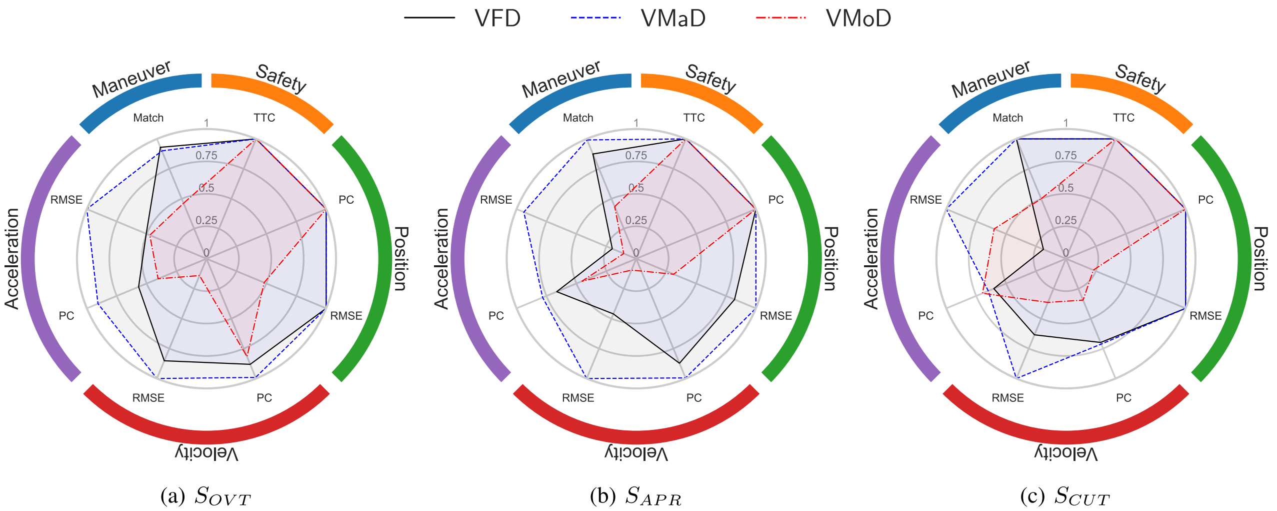
The results for the utilized simulation setup are depicted in Fig. 12. A larger occupied area within the kiviat chart is associated with higher credibility. There are no major differences in quality between the three scenarios presented. The maneuver-based extraction of simulation scenarios from the measurement data works very well. A perfect mapping is not possible, because smoothing and linearization is performed during the classification into maneuvers. Short-term and fine-granular deviations of the physical parameters are lost in the process. In the VFD the coarse-granular quantities such as the safety and the maneuvers are reproduced accurately. However, it is noticeable in the physical quantities that the quality decreases from position to velocity to acceleration. The results suggest that better tuning may be possible between the ADF and the vehicle dynamics model. Afterwards, one could re-trigger the evaluation chain and check if an improvement has occurred. As expected, VMoD performs worst in the credibility assessment.

The normalized OSPA as combined metric of several input data correlates with the areas in the kiviat diagrams for VFD, VMaD as well as for VMoD (Table 4). The lower values compared to the areas in the radar chart are due to criticality and maneuver measures, which are not an input of the OSPA metric, but perform very well in each of the three scenarios. Consequently, a credibility statement relative to other simulation setups or thresholds is possible with the introduction of normalized metrics.

VI. DISCUSSION

Thanks to measurement data from a prototype automated vehicle, an end-to-end validation loop can be closed across the entire co-simulation system. Therein, the response of the ADF to the virtual events is contrasted with the response of the ADF in the real test drive. The quality of the ADF is not in the scope of the work, merely the demonstration of a comparable behavior in the simulation.

An important aspect of reproducing the test drive in the simulation is the generation of the OSC file and the associated maneuver drive. The ADF interacts with the traffic participants in the simulation. If these behave with a too big deviation from the test drive data, this can also severely impair the reaction of the ADF in simulation, because there is



(a) S_{OVT} (b) S_{APR} (c) S_{CUT}

FIGURE 12. **Multivariate credibility results after normalization.**

an interplay between ego and external vehicles. Particularly critical points in time are the threshold zones between deviating decision making, which is reflected in particular in the increased error in S_{APR}, where the ego vehicle constantly faces a minimum distance threshold to the vehicle it follows. A trajectory-based generation of the scenarios could certainly produce better results than the maneuver-based one in terms of reproducing the traffic scenario. However, accuracy and flexibility must be weighed here. Only maneuver-based scenarios can be reasonably altered by parameter variation, allowing for greater test space coverage. Furthermore, it is only possible in maneuver-based scenarios that interaction between the automated ego vehicle and other traffic participants can take place. In trajectory-based scenarios, the trajectory is rigorously followed by the traffic participant. This is considered to be an exclusion criterion for tarjectory-based scenarios, since they have no relevance, at least for the presented use case.

Through the evaluation approach from coarse to fine, profound insights for the state of simulation for the safeguarding of ADF can be gained, where until today it is unclear which simulation quality is sufficient. The visual evaluation of the video footage of the real and the virtual function drive shows that the drives were very similar and it was difficult to detect a difference. Lateral and longitudinal maneuver evaluation further confirm this result. In the RMSE, this has an effect of a trajectory deviation of a few single-digit meters. The threshold of a valid to an invalid simulation can be assumed in approximately in this range as the velocities driven in the scenarios are greater than $20[m/s]$. A detailed determination of these thresholds are depended on the requirements of the test case. A distinction must also be made between lateral offset, where a deviation is more severe, and longitudinal offset, where it is less severe. The same is true for distant objects, for which relative positioning to the ego vehicle is less important, and for close objects, for which relative positioning is central to the decision-making of the ADF.

As the results in Section V illustrate, the resimulation of the three scenarios considered yields fair results both qualitatively and quantitatively. Nevertheless, the presented procedure for the methodology shows three limitations respectively uncertainties which came to the authors notice during the case study.

As an assumption ideal sensor models are used, which exactly detect the vehicles environment as it is modeled within ODR for the static and within OSC for dynamic entities. Therefore, possible uncertainties originated from noisy sensor signals of the perception system are not reproduced in simulation as well as uncertainties of state, class and existence of the objects and weather effects. This assumptions is necessary to limit the occurring deviations to the steps scenario extraction, simulation environment and simulation execution. However, high fidelity sensor models, like phenomenological or physical ones can be integrated in the procedure as well.

As the presented approach is data driven, the credibility of the resimulated scenarios compared to the real test drive strongly depends on the quality of the measurement data in terms of accuracy and availability. As the information for the modelling of the surrounding traffic is based on the perception system of the ego vehicle and is measured relatively to ego. Consequently, the worse the sensor equipment and the object recognition, the less accurate are the input data for extracting the behavior of the other traffic participants. This results in propagating errors and increases the uncertainty in modeling of the surrounding traffic within the OSC. Therefore, it is reasonable that higher deviations are detected in the validation step for these objects compared to the ego vehicle. This point is not the case during the conducted case study but should generally be taken into account during validation comparison.

The presented simulations are conducted with VTD as tool for driving simulation. However, the approach from Section IV is generally transferable and applicable with every state of the art simulation tooling which supports ASAM standards and guarantees strong determinism. Nevertheless, some simulators may interpret OSC and ODR files slightly different internally. This circumstance provokes deviations by using a different tool set. Therefore, metadata are needed for the simulation tools which provide information about the interpretation of specific data as well as for the ASAM standards as simulation input in order to perform a tool independently reliable analysis for validation purposes.

VII. CONCLUSION

The approval of ADFs is a huge challenge in the automotive industry and research. In order to validate an ADF, simulation- and scenario-based approaches are rising in popularity. Because of the immense test space, it is known that the validation without the use of computer simulations might not be feasible or practical. The question of the credibility of the scenario-based SIL simulation arises. An ADF is designed and developed for functioning in the real world, but the testing is intended to be done mostly virtually. If the same ADF behaves differently in the simulation than in the real world, the results in the simulation have no significance.

In this work a methodology for assessing simulation scenarios and quantifying their credibility is presented: A real test drive is performed with an automated vehicle and the recorded data are elaborated in order to resimulate the occurred scenarios. In the simulation the ADF is confronted with the extracted scenario in a virtual environment of the road network. If the behavior of the ADF in the physical and in the virtual world is exactly or approximately the same, the simulation is considered to be credible. The simulation setup can only be tested and calibrated with ground truth data, i.e., only against scenarios driven in reality.

The method shows promising results. The authors are able to show similar behavior in five different categories and metrics: maneuvers, criticality, correlation, RMSE and OSPA. In order to have a reference the simulation results are normalized and compared to threshold with is dependent on the requirements of the test case.

As future work, this analysis should be performed on a larger scale in order to understand which type of scenarios are more credible in the simulation compared to others. In this way, it could be defined in which component the simulation has to improve and what type of scenarios should be tested more accurately in the real world. In this work the credibility is quantified with not further defined thresholds. However, as future work it is intended to find thresholds for different quality steps of the credibility, which are derived from the requirement specifications. As described in Section VI the quality of the measurement data, especially of the environmental sensors, is responsible for the extraction and reconstruction of the behavior of the traffic participants in the OSC. An idea for improving and validating the data could be to perform drives with multiple vehicles that can see each other. In this way, non ideal sensors can also be added in the simulation and perform the credibility analysis also for the sensor set. Lastly, it would be interesting to use the proposed method on different state-of-the-art simulation tools in order to establish and understand the differences in quality.

REFERENCES

[1] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Transp. Safety*, vol. 4, no. 1, pp. 15–24, 2016.

[2] W. Wachenfeld and H. Winner, "The release of autonomous vehicles," in *Autonomous Driving*. Heidelberg, Germany: Springer, 2016, pp. 425–449.

[3] H. Winner, K. Lemmer, T. Form, and J. Mazzega, "PEGASUS—First steps for the safe introduction of automated driving," in *Road Vehicle Automation 5* (Lecture Notes in Mobility). Cham, Switzerland: Springer Int., 2019, pp. 185–195.

[4] A. Ngo, M. P. Bauer, and M. Resch, "A multi-layered approach for measuring the simulation-to-reality gap of radar perception for autonomous driving," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, 2021, pp. 4008–4014.

[5] K. Groh, S. Wagner, T. Kuehbeck, and A. Knoll, "Simulation and its contribution to evaluate highly automated driving functions," *SAE Int. J. Adv. Current Pract. Mobility*, vol. 1, no. 2, pp. 539–549, 2019.

[6] D. Notz *et al.*, "Methods for improving the accuracy of the virtual assessment of autonomous driving," in *Proc. IEEE Int. Conf. Connected Veh. Expo (ICCVE)*, 2019, pp. 1–6.

[7] S. Wagner, K. Groh, T. Kühbeck, and A. Knoll, "Towards cross-verification and use of simulation in the assessment of automated driving," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2019, pp. 1589–1596.

[8] A. Höfer and M. Herrmann, *Scenario-Based Approach for Developing ADAS and Automated Driving Functions*. Wiesbaden, Germany: Springer-Verlag, 2017, pp. 215–225.

[9] C. Sippl, F. Bock, C. Lauer, A. Heinz, T. Neumayer, and R. German, "Scenario-based systems engineering: An approach towards automated driving function development," in *Proc. IEEE Int. Syst. Conf. (SysCon)*, Apr. 2019, pp. 1–8.

[10] G. Bagschik, T. Menzel, C. Körner, and M. Maurer, "Wissensbasierte szenariengenerierung für betriebsszenarien auf deutschen autobahnen," in *Proc. Workshop Fahrerassistenzsysteme Automatisiertes Fahren*, vol. 12, 2018, p. 14.

[11] L. Wang *et al.*, "Multi-functional open-source simulation platform for development and functional validation of ADAS and automated driving," in *Fahrerassistenzsysteme 2016*, R. Isermann, Ed. Wiesbaden, Germany: Springer-Verlag, 2018, pp. 135–148.

[12] F. A. Schiegg, J. Krost, S. Jesenski, and J. Frye, "A novel simulation framework for the design and testing of advanced driver assistance systems," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, 2019, pp. 1–6.

[13] H. Abdellatif and C. Gnandt, "Use of simulation for the homologation of automated driving functions," *ATZelectronics Worldwide*, vol. 14, no. 12, 2019, pp. 68–71.

[14] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, "Defining and substantiating the terms scene, situation, and scenario for automated driving," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, 2015, pp. 982–988.

[15] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for thedevelopment of automated vehicles," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2018, pp. 1813–1820.

[16] "ASAM SIM:GUIDE—Standardization for highly automated driving," ASAM e.V., Hoehenkirchen, Germany, Rep., 2021. [Online]. Available: https://www.asam.net/asam-guide-simulation/

[17] R. Roe, "Standard for models and simulations," Nat. Aeronaut. Space Admin., Washington, DC, USA, Rep. NASA-STD-7009A, 2016.

[18] F. Liu, M. Yang, and Z. Wang, "Study on simulation credibility metrics," in *Proc. Winter Simulat. Conf.*, 2010, pp. 166–183.

[19] R. Rabeau, "Credibility in modeling and simulation," in *Simulation and Modeling of Systems of Systems*. Chichester, U.K.: Wiley, 2013, ch. 3, pp. 99–157.

[20] A. Erdogan *et al.*, "Real-world maneuver extraction for autonomous vehicle validation: A comparative study," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2019, pp. 267–272.

[21] *Mercedes-Benz Präsentiert in Genf Limousine und Coupé Der Neuen E-Klasse*, document, Daimler AG, Stuttgart, Germany, 2009. [Online]. Available: https://www.presseportal.de/download/document/115362-pi-mb-genf-2009-d.pdf

[22] F. Montanari, R. German, and A. Djanatliev, "Pattern recognition for driving scenario detection in real driving data," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 590–597.

[23] F. Montanari, H. Ren, and A. Djanatliev, "Scenario detection in unlabeled real driving data with a rule-based state machine supported by a recurrent neural network," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, 2021, pp. 1–5.

[24] S. P. Hoogendoorn and P. H. L. Bovy, "State-of-the-art of vehicular traffic flow modelling," *Proc. Inst. Mech. Eng. I, J. Syst. Control Eng.*, vol. 215, no. 4, p. 283–303, 2001.

[25] R. L. Bucs, R. Leupers, and G. Ascheid, "Multi-scale multi-domain co-simulation for rapid ADAS prototyping," in *Proc. IEEE Asia Pac. Conf. Circuits Syst. (APCCAS)*, Oct. 2018, p. 532–535.

[26] M. Paulweber, "Validation of highly automated safe and secure systems," in *Automated Driving*. Cham, Switzerland: Springer Int., 2017, pp. 437–450.

[27] J. Hertzberg, K. Lingemann, and A. Nüchter, "Roboterkontrollarchitekturen," in *Mobile Roboter*. Heidelberg, Germany: Springer, 2012, pp. 317–333.

[28] W. Baron, C. Sippl, K.-S. Hielscher, and R. German, "Repeatable simulation for highly automated driving development and testing," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–7.

[29] W. Baron, A. Arestova, C. Sippl, K.-S. Hielscher, and R. German, "LETT: An execution model for distributed real-time systems," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2021, p. 1–7.

[30] "Pegasus method," Deutsches Zentrum für Luft- und Raumfahrt e.V., Institut Verkehrssystemtechnik, Braunschweig, Germany, Research Rep., 2019. [Online]. Available: https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf

[31] P. Koopman and M. Wagner, "Toward a framework for highly automated vehicle safety validation," Warrendale, PA, USA, SAE Int., Technical Paper, 2018.

[32] *Proposal for a New UN Regulation on Uniform Provisions Concerning the Approval of Vehicles With Regards to Automated Lane Keeping System*, document ECE/TRANS/WP.29/2020/81, United Nat. Econ. Commiss. Europe (UNECE), Geneva, Switzerland, 2020.

[33] "Road vehicles—Functional safety," Int. Org. Stand., Geneva, Switzerland, Rep. ISO 26262-2:2018, 2018.

[34] R. Sargent, "Verification and validation of simulation models," in *Proc. Winter Simulat. Conf.*, 2010, pp. 166–183.

[35] R. Sargent and D. Goldsman, "Use of the internal statistical procedure for simulation model validation," in *Proc. Winter Simulat. Conf.*, 2015, pp. 60–72.

[36] S. Riedmaier, D. Schneider, D. Watzenig, F. Diermeyer, and B. Schick, "Model validation and scenario selection for virtual-based homologation of automated vehicles," *Appl. Sci.*, vol. 11, no. 1, p. 35, 2021.

[37] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," *Accid. Anal. Prevent.*, vol. 163, Dec. 2021, Art. no. 106454.

[38] F. Montanari, C. Stadler, J. Sichermann, R. German, and A. Djanatliev, "Maneuver-based resimulation of driving scenarios based on real driving data," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2021, pp. 1124–1131.

[39] D. Schuhmacher, B. Vo, and B. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 996–1005, Aug. 2008.

[40] B. Ristic, B. Vo, D. Clark, and B. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3452–3457, Jul. 2011.

[41] C. Stadler, K. Rauner, R. German, and A. Djanatliev, "Simulation-based parameter identification for accuracy defnitions in virtual environment models for validation of automated driving," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2021, pp. 1138–1143.

[42] A. R. Mamdoohi, M. Fallah Zavareh, C. Hydén, and T. Nordfjærn, "Comparative Analysis of Safety Performance Indicators Based on Inductive Loop Detector Data," *Promet Traffic Transp.*, vol. 26, no. 2, pp. 139–149, Apr. 2014.

[43] B. Huber, S. Herzog, C. Sippl, R. German, and A. Djanatliev, "Evaluation of virtual traffic situations for testing automated driving functions based on multidimensional criticality analysis," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, 2020, pp. 1–7.

[44] D. Freedman, R. Pisani, and R. Purves, *Statistics (International Student Edition)*, 4th ed. New York, NY, USA: Norton Company, 2007.

[45] C. Beisbart and N. J. Saam, *Computer Simulation Validation. Fundamental Concepts, Methodological Frameworks, Philosophical Perspectives*. Cham, Switzerland: Springer Nat., 2019.

[46] K. von Neumann-Cosel, "Virtual test drive," Ph.D. dissertation, Dept. Fakultät Informatik, Technische Universität München, Munich, Germany, 2014.

CHRISTOPH STADLER received the B.Sc. degree in engineering science in 2015, and the M.Sc. degree in mechanical engineering from the Technical University of Munich, Germany, in 2018. He is currently pursuing the Ph.D. degree in a collaboration between the AUDI AG and the Chair of Computer Networks and Communication Systems of the Friedrich-Alexander University Erlangen-Nuremberg, Germany. His research interests are focused on quality requirements for virtual test fields and credibility of simulation results for automated driving.

FRANCESCO MONTANARI received the B.Sc. and M.Sc. degrees in electrical and computer engineering from the Technical University of Munich, Germany, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree in a collaboration between the AUDI AG and the Computer Networks and Communication Systems Chair of the Computer Science Department, University of Erlangen-Nuremberg, Germany. His research interests are focused on driving data analysis, scenario identification, and resimulation of scenarios.

WOJCIECH BARON received the B.Sc. and M.Sc. degrees in information and communication technology from the University of Erlangen-Nuremberg, Germany, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. He collaborates with AUDI AG in diverse projects that evaluate synchronization mechanisms and real-time capabilities in distributed simulation systems in the context of automated driving.

CHRISTOPH SIPPL received the B.Sc. degree in medical information technology from Ostbayerische Technische Hochschule Regensburg in 2012, and the M.Sc. and Ph.D. (Dr.-Ing.) degrees in computer science from the Engineering Faculty, University of Erlangen-Nuremberg in 2014 and 2020, respectively. His current research interests are focused on scenario-based development and virtual validation methods for automated driving.

ANATOLI DJANATLIEV received the M.Sc. and Ph.D. (Dr.-Ing.) degrees in computer science (Dipl.-Inf. Univ.) from the Engineering Faculty, University of Erlangen-Nuremberg in 2015 and 2008, respectively. He is the Head of the Connected Mobility Group with the Chair of Computer Networks and Communication Systems, FAU. His current research interests include various topics on simulation and modeling. Major application areas are simulation of vehicular networks, innovative aspects of connected mobility, and future mobility services.