Requirements-driven Test Generation for Autonomous Vehicles with Machine Learning Components*

Cumhur Erkan Tuncali[1], Georgios Fainekos[1], Danil Prokhorov[2], Hisahiro Ito[2], and James Kapinski[2]

***Abstract*—Autonomous vehicles are complex systems that are challenging to test and debug. A requirements-driven approach to the development process can decrease the resources required to design and test these systems, while simultaneously increasing the reliability. We present a testing framework that uses signal temporal logic (STL), which is a precise and unambiguous requirements language. Our framework evaluates test cases against the STL formulae and additionally uses the requirements to automatically identify cases that fail to satisfy the requirements. One of the key features of our tool is the support for machine learning (ML) components in the system design, such as deep neural networks. The framework allows evaluation of the control algorithms, including the ML components, and it also includes models of CCD camera, lidar, and radar sensors, as well as the vehicle environment. We use multiple methods to generate test cases, including covering arrays, which is an efficient method to search discrete variable spaces. The resulting test cases can be used to debug the controller design by identifying controller behaviors that do not satisfy requirements. The test cases can also enhance the testing phase of development by identifying critical corner cases that correspond to the limits of the system's allowed behaviors. We present three STL requirements for an autonomous vehicle system, which capture both component-level and system-level behaviors. Additionally, we present three driving scenarios and demonstrate how our requirements-driven testing framework can be used to identify critical system behaviors, which can be used to support the development process.**

I. Introduction

Autonomous driving systems are in a stage of rapid research and development spanning a broad range of maturity from simulations to on-road testing and deployment. They are expected to have a significant impact on the vehicle market and the broader economy and society in the future.

Testing of highly automated and autonomous driving systems is also an area of active research. Both governmental and non-governmental organizations are grappling with the unique requirements of these new, highly complex systems, as they have to operate safely and reliably in diverse driving environments. Government and industry sponsored partnerships have produced a number of guiding documents and clarifications, such as NHTSA [1], SAE [2], CAMP [3], NCAP [4], PEGASUS [5]. The research community has also been contributing to the development of methodologies for testing automated driving systems.

Stellet et al. [6] surveyed existing approaches to testing such as simulation-only, X-in-the-loop and augmented reality approaches, as well as test criteria and metrics (see also [7]). Koopman and Wagner identified challenges of testing and proposed potential solutions, such as fault injection, as a way to perform more efficient edge case testing [8]. The publications [9] and [10] provide in-depth discussions on the challenges of safety validation for autonomous vehicles, arguing that virtual testing should be the main target for both methodological and economic reasons.

No universally agreed upon testing or verification methods have yet arisen for autonomous driving systems. One reason is that the current autonomous systems architectures usually include some Machine Learning (ML) components, such as Deep Neural Networks (DNNs), which are notoriously difficult to test and verify. We present a framework for Simulation-based Adversarial Testing of Autonomous Vehicles (Sim-ATAV), which can be used to check closed-loop properties of autonomous driving systems that include ML components. We describe a testing methodology, based on a test case generation method, called covering arrays, and requirement falsification methods to automatically identify problematic test scenarios. The resulting framework can be used to increase the reliability of autonomous driving systems.

Autonomous driving system designs often use ML components such as DNNs to classify objects within CCD images and to determine their positions relative to the vehicle, a process known as *object detection and classification* [11]. Other designs use Neural Networks (NNs) to perform *end-to-end* control of the vehicle, meaning that the NN takes in the image data and outputs actuator commands, without explicitly performing an intermediate object detection step [12], [13], [14]. Still other approaches use end-to-end learning to do intermediate decisions like risk assessment [15].

ML system components are problematic from an analysis perspective, as it is difficult or impossible to characterize all of the behaviors of these components under all circumstances. One reason is that the complexity of these systems can be very high in terms of the number of parameters. For example, AlexNet [16], a pre-trained DNN that is used for classification of CCD images, has 60 million parameters. Another reason for the difficulty in characterizing behaviors of ML components is that the parameters are learned based on training data. In other words, characterizing ML behaviors is, in some ways, as difficult as the task of characterizing the training data. Again using the AlexNet example, the number of training images used was 1.2 million. While a strength of DNNs is their ability to generalize from training data; the challenge for analysis is that we do not understand well how they generalize for all possible cases.

There has been significant interest recently on verification and testing for ML components (see Sec. II). For example, adversarial testing approaches seek to identify perturbations in image data that result in misclassifications. By contrast,

*This work was partially funded by NSF awards CNS 1446730, 1350420

[1]School of Computing, Informatics & Decision Systems Engineering, Arizona State University, USA etuncali, fainekos@asu.edu

[2]Toyota Technical Center, Ann Arbor MI, USA danil.prokhorov, hisahiro.ito@toyota.com, jim.kapinski@toyota.com

our work focuses on methods to determine perturbations in the configuration of a testing scenario, meaning that we seek to find scenarios that lead to unexpected behaviors, such as misclassifications and ultimately vehicle collisions. The framework that we present allows this type of testing in a virtual environment. By utilizing advanced 3D models and image rendering tools, such as the ones used in game engines or film studios, the gap between testing in a virtual environment and the real world can be minimized.

Most of the previous work to test and verify systems with ML components focuses only on the ML components themselves, without consideration of the closed-loop behavior of the system. For autonomous driving applications, we remark that the ultimate goal is to evaluate the closed-loop system performance, and hence, any testing methods used to evaluate such systems should support this goal.

The closed-loop nature of a typical autonomous driving system can be described as follows. A *perception* system processes data gathered from various sensing devices, such as cameras, lidar, and radar. The output of the perception system is an estimation of the principal (*ego*) vehicle's position with respect to external obstacles (e.g., other vehicles, called *agent* vehicles, and pedestrians). A *path planning* algorithm uses the output of the perception system to produce a short-term plan for how the ego vehicle should behave. A *tracking controller* then takes the output of the path planner and produces actuation outputs, such as accelerator, braking, and steering commands. The actuation commands affect the vehicle's interaction with the environment. The iterative process of sensing, processing, and actuating is what we refer to as closed-loop behavior.

An earlier version of this work appeared in [17]. The contributions of that work can be summarized as follows. In [17], we provided a new algorithm to perform falsification of formal requirements for an autonomous vehicle in a closed-loop with the perception system, which includes an efficient means of searching over discrete and continuous parameter spaces. The method represents a new way to do adversarial testing in scenario *configuration space*, as opposed to the usual method, which considers adversaries in *image space*. Additionally, we demonstrated a new way to characterize problems with perception systems in *configuration space*. Lastly, we extended the software testing theory of covering arrays to closed-loop Cyber-Physical System (CPS) applications that have embedded ML algorithms.

The present paper provides the following contributions that are in addition to those from [17].

- We add models of lidar and radar sensors and include sensor fusion algorithms, and we demonstrate how the requirements-based testing framework we propose can be used to automate the search for specific types of fault cases involving sensor interactions.
- We provide requirements for both component-level and system-level behaviors, and we show how to automate the identification of behaviors where component-level failures lead to system-level failures. An example of the kind of analysis this allows is automatically finding cases where a sensor failure leads to a collision case.
- We include a model of agent *visibility* to various sensors and include this notion in the requirements that we consider. This provides a way to reason about how the system should behave, based on whether agents are or are not visible, including the ability to reason about the temporal aspects of agent visibility. For example, we can use this feature to test the requirement that within 1 second after an agent becomes visible to the lidar sensor, the perception system should correctly classify the agent. This allows us to automate the search for behaviors related to temporal aspects of sensor behaviors in the context of a realistic driving scenario.
- We demonstrate the ability to falsify properties by adversarially searching over agent trajectories. This permits the use of our requirements-driven search-based approach over a broad class of agent behaviors, which allows us to automatically identify corner cases that are difficult to find using traditional simulation-based techniques.

II. Related work

Testing and evaluation methods for Autonomous[1] Vehicles (AVs) could be categorized into three major classes: (1) model based, (2) data-driven, and (3) scenario based. Scenario-based approaches utilize accident reports and driving conditions that are easily identifiable as challenging, producing specific test scenarios to be executed either in the real world or in a simulation environment. For example, Euro NCAP [4] and DOT [18] provide such scenarios. Data-driven approaches, on the other hand, typically utilize driving data [19] to generate probabilistic models of human drivers. Such models are then used for risk assessment and rare event sampling for AV algorithms under specific driving scenarios [20].

The aforementioned testing methods are important and necessary before AV deployment, but they cannot help with design exploration and automated fault detection at early development stages. Such problems are addressed by model-based verification [21], [22], model based test generation [23], [24], [25], [26], [27], [28], or a combination thereof [29], [30]. It is important to also highlight that these methods typically ignore or use simple models to abstract away proximity sensors and, especially, the vision systems. However, ignoring sensors or using simplified sensing models may be a dangerously simplifying assumption since it ignores the complex interactions between the dynamics of the vehicle and the sensors. For example, the effective sensing range of a sensor platform mounted on the roof of a vehicle is affected when the vehicle makes hard turns.

In addition, vision-based perception systems have become an integral component of the sensor platform of AVs, and in many cases, they constitute the only perception system. Currently, the winning algorithmic technology for image processing systems is utilizing Deep Neural Networks (DNN). For instance, by 2011, the DNN architecture proposed in [31] was already capable of classifying pre-segmented images of traffic

[1]We utilize the more general term "autonomous" as opposed to a more restricted "automated" since our methods could potentially apply to all levels of autonomy.

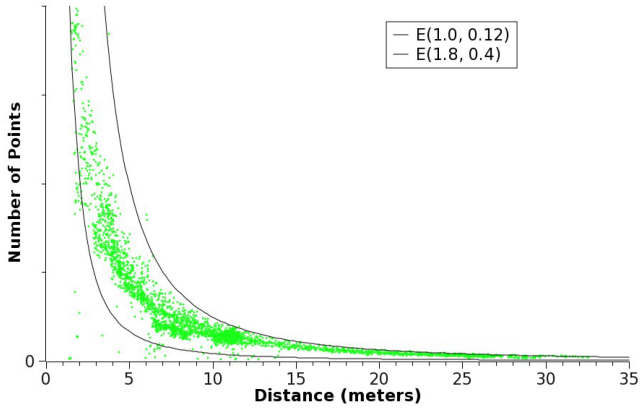


Fig. 1: Lidar reflection points versus distance to various pedestrian and pedestrian-like targets. Lines of expected number of points $E(h, w)$ are also shown.

signs with better accuracy than humans (99.46% vs 99.22%). Since then, there has been substantial progress with DNNs performing both segmentation and classification [32], [33]. Yet, in spite of the multiple impressive results using DNN, it is still also easy to devise methods that can produce (so-called adversarial) images that will fool them [34], [35], [36].

The latter (negative) result raises two important questions: (1) can we still generate adversarial inputs for DNN when we manipulate the physical properties and trajectories of the objects in the environment of the AV, and (2) how does the DNN accuracy affect the system level properties of an AV, that is, its functional safety? Exhaustive verification methods for NN in the loop are still in their infancy [37], and they cannot handle AV with DNN components in the loop. To address the two questions above, several model-based test generation methods have been proposed [38], [39], [40], [17]. The procedure described in [38], [39] analyzes the performance of the perception system using static images to identify candidate counterexamples, which are then checked using simulations of the closed-loop system to determine whether the AV exhibits unsafe behaviors. On the other hand, [40], [17] develop methods that directly search for unsafe behaviors of the closed-loop system by defining a cost function on the closed-loop behaviors. The differences between [40] and [17] are primarily on the search methods, the simulation environments, and the AVs, with [17] providing a more efficient method for combinatorial search.

In this extended version of [17], we take the system-level adversarial test generation methods for AV one step further. We demonstrate that our framework [17] can be extended for test generation for AV with multi-sensor systems as opposed to vision-only perception systems. Moreover, we demonstrate the importance and effectiveness of test generation methods guided by system-level requirements as well as perception-level requirements.

Using our framework, we can formalize and test against requirements on the sensor performance, in the context of a driving scenario. For example, the lidar's point cloud density drops significantly with the distance to the target object, for example, a pedestrian; Figure 1 illustrates this point with experimental results showing lidar data point density as a function of object distance. Similar to this aspect of lidar behavior, the pixel count of a CCD camera would also decrease dramatically with the distance if it were to be used for pedestrian detection, since the area of an observed object decreases as the square of the distance to the object. This may complicate testing for long-range observation conditions. Our framework supports testing these aspects of sensor performance.

III. Preliminaries

This section presents the setting used to describe the testing procedures performed with our framework. The purpose of our framework is to provide a mechanism to test, evaluate, and improve on an autonomous driving system design. To do this, we use a simulation environment that incorporates models of a vehicle (called the *ego* vehicle), a perception system, which is used to estimate the state of the vehicle with respect to other objects in its environment, a controller, which makes decisions about how the vehicle will behave, and the environment in which the ego vehicle is deployed. The environment model contains representations of a wide variety of objects that can interact with the ego vehicle, including roads, buildings, pedestrians, and other vehicles (called *agent* vehicles). The behaviors of the system are determined by the evolution of the model states over time, which we compute using a *simulator*.

Formally, the framework implements a model of the system, which is a tuple $M = (\mathcal{X}, \mathcal{U}, \mathcal{P}, sim)$, where $\mathcal{X}$ is a set of system states, $\mathcal{U}$ is a set of inputs, and sim is a simulation function $sim : \mathcal{X} \times \mathcal{U} \times \mathcal{P} \times \mathcal{T} \to \mathcal{X}$, where $\mathcal{T}$ is a discrete set of sample times $t_0, t_1, \ldots, t_N$, with $t_i < t_{i+1}$. $\mathcal{P} = \mathcal{W} \times \mathcal{V}$ is a combination of continuous-valued and discrete-valued parameters, where $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_W$ and each $\mathcal{W}_i \subseteq \mathbb{R}$, and $\mathcal{V} = \mathcal{V}_1 \times \cdots \times \mathcal{V}_V$ and each $\mathcal{V}_i$ is some finite domain, such as Boolean or a finite list of agent car colors.

Given $\mathbf{x} \in \mathcal{X}$, $\hat{\mathbf{x}} = sim(\mathbf{x}, \mathbf{u}, \mathbf{p}, t)$ is the state reached starting from state $\mathbf{x}$ after time $t \in \mathcal{T}$ under input $\mathbf{u} \in \mathcal{U}$ and parameter value $\mathbf{p} \in \mathcal{P}$. We call a sequence

$$U = (\mathbf{u}_0, t_0)(\mathbf{u}_1, t_1) \cdots (\mathbf{u}_N, t_N),$$

where each $\mathbf{u}_i \in \mathcal{U}$ and $t_i \in \mathcal{T}$, an input trace of M. Given a model M, an input trace of M, U, and a $\mathbf{p} \in \mathcal{P}$, a simulation trace of M under input U and parameters $\mathbf{p}$ is a sequence

$$T = (\mathbf{x}_0, \mathbf{u}_0, t_0)(\mathbf{x}_1, \mathbf{u}_1, t_1) \cdots (\mathbf{x}_N, \mathbf{u}_N, t_N),$$

where $sim(\mathbf{x}_{i-1}, \mathbf{u}_{i-1}, \mathbf{p}, t_{i-1}) = \mathbf{x}_i$ for each $1 \leq i \leq N$. For a given simulation trace T, we call $X = (\mathbf{x}_0, t_0)(\mathbf{x}_1, t_1) \cdots (\mathbf{x}_N, t_N)$ the state trace. We denote the set of all simulation traces of M by $\mathcal{L}(M)$.

A. Signal Temporal Logic

Signal Temporal Logic (STL) was introduced as an extension to Metric Temporal Logic (MTL) to reason about real-time properties of signals (simulation traces) (for an overview see [41]). STL formulae are built over predicates on the variables of a signal using combinations of Boolean and temporal operators. The temporal operators include *eventually* ($\Diamond_{\mathcal{I}}$), *always* ($\Box_{\mathcal{I}}$) and *until* ($\mathcal{U}_{\mathcal{I}}$), where $\mathcal{I}$ is a time interval

that encodes timing constraints. The boolean operators include *conjunction* $\wedge$, *disjunction* $\vee$, *negation* $\neg$, and *implies* $\implies$.

In this work, we interpret STL formulas over the observable simulation traces of a given system. STL specifications can describe the usual properties of interest in system design such as (bounded time) **reachability**, e.g., *between time 1.2 and 5 (not including),* $\mathbf{x}$ *should drop below* -10: $\Diamond_{[1.2,5)}\ (\mathbf{x} \leq -10)$, and **safety**, e.g., *after time 2,* $\mathbf{x}$ *should always be greater than* 10: $\Box_{[2,+\infty)}(\mathbf{x} \geq 10)$. STL can capture sequences of events, e.g., $\Diamond_{\mathcal{I}_1}(\pi_1 \wedge \Diamond_{\mathcal{I}_2}(\pi_2 \wedge \Diamond_{\mathcal{I}_3}\pi_3))$, which states that π_1 should become true at some time in the time interval $\mathcal{I}_1$ followed by at least one event satisfying π_2 within time $\mathcal{I}_1 \oplus \mathcal{I}_2$, followed in turn by a satisfying event for π_3 in the time interval $\mathcal{I}_1 \oplus \mathcal{I}_2 \oplus \mathcal{I}_3$. Here, $\oplus$ is the Minkowski sum and π_i are predicates over signal variables, e.g., $\pi_1 \equiv x_1 + 3x_2 \leq 4.5$. Another example of expressible requirements in STL are infinite occurring behaviors, e.g., **reactive** behaviors : $\Box(\pi_1 \rightarrow \Diamond_{[0,2]}\pi_2)$, which states that whenever π_1 is satisfied, then within 2 time units π_2 should also become true.

Informally speaking, we allow predicate expressions to capture arbitrary constraints over the state variables, inputs, and parameters of the system. More formally, we assume that predicates π are expressions built using the grammar $\pi ::= f(\mathbf{x}, \mathbf{u}, \mathbf{p}) \geq c \mid \neg\pi_1 \mid (\pi) \mid \pi_1 \vee \pi_2 \mid \pi_1 \wedge \pi_2$, where f is a function and c is a constant in $\mathbb{R}$. In other words, each predicate π represents a subset in the space $\mathcal{X} \times \mathcal{U} \times \mathcal{P}$. In the following, we represent the set that corresponds to the predicate π using the notation $\mathcal{O}(\pi)$. For example, if $\pi = (x_1 \leq -10) \vee (x_1 + x_2 \geq 10)$ and we represent by x_i the i-th component of the vector $\mathbf{x}$, then $\mathcal{O}(\pi) = (\infty, -10] \times \mathbb{R} \cup \{x \in \mathbb{R}^2 \mid x_1 + x_2 \geq 10\}$.

Definition 1 (STL Syntax): Assume Π is the set of predicates and $\mathcal{I}$ is any non-empty connected interval of $\mathbb{R}_{\geq 0}$. The set of all well-formed STL formulas is inductively defined as $\varphi ::= \top \mid \pi \mid \neg\phi \mid \phi_1 \vee \phi_2 \mid \bigcirc \phi \mid \phi_1 U_{\mathcal{I}} \phi_2$, where π is a predicate, $\top$ is *true*, $\bigcirc$ is Next, and $U_{\mathcal{I}}$ is the Until operator.

In this work, we will be using discrete time semantics of STL since we would like to be able to reason about the timing of samples and define events as falling or raising Boolean values, using the next time operator ($\bigcirc$). For example, the formula $\pi \wedge \neg\pi$ expresses a falling value event. For STL formulas ψ, ϕ, we define $\psi \wedge \phi \equiv \neg(\neg\psi \vee \neg\phi)$, $\bot \equiv \neg\top$ (False), $\psi \rightarrow \phi \equiv \neg\psi \vee \phi$ (ψ Implies ϕ), $\Diamond_I \psi \equiv \top U_I \psi$ (Eventually ψ), $\Box_I \psi \equiv \neg\Diamond_I \neg\psi$ (Always ψ), and $\psi R_I \phi \equiv \neg(\neg\psi U_I \neg\phi)$ (ψ Releases ϕ), using syntactic manipulation.

In our previous work [42], we proposed robust semantics for STL formulas. Robust semantics (or robustness metrics) provide a real-valued measure of satisfaction of a formula by a trace, in contrast to the Boolean semantics that just provide a *true* or *false* valuation. In more detail, given a trace T of the system, its robustness w.r.t. a temporal property φ, denoted $[\![\varphi]\!]_{\mathbf{d}}(T)$ yields a positive value if T satisfies φ and a negative value otherwise. Moreover, if the trace T satisfies the specification ϕ, then the robust semantics evaluate to the radius of a neighborhood such that any other trace that remains within that neighborhood also satisfies the same specification. The same holds for traces that do not satisfy ϕ.

Definition 2 (STL Robust Semantics): Given a metric $\mathbf{d}$, trace T, and $\mathcal{O} : \Pi \rightarrow 2^{\mathcal{X} \times \mathcal{U} \times \mathcal{P}}$, the robust semantics of any formula ϕ w.r.t T at time instance $i \in N$ is defined as:

$$
\begin{aligned}
[\![\top]\!]_{\mathbf{d}}(T, i) &:= +\infty \\
[\![\pi]\!]_{\mathbf{d}}(T, i) &:= \begin{cases} -\inf\{\mathbf{d}((\mathbf{x}_i, \mathbf{u}_i, \mathbf{p}_i), \mathbf{y}) \mid \mathbf{y} \in \mathcal{O}(\pi)\} \\ \quad \text{if } (\mathbf{x}_i, \mathbf{u}_i, \mathbf{p}_i) \notin \mathcal{O}(\pi) \\ \inf\{\mathbf{d}((\mathbf{x}_i, \mathbf{u}_i, \mathbf{p}_i), \mathbf{y}) \mid \mathbf{y} \in \overline{\mathcal{O}(\pi)}\} \\ \quad \text{if } (\mathbf{x}_i, \mathbf{u}_i, \mathbf{p}_i) \in \mathcal{O}(\pi) \end{cases} \\
[\![\neg\phi]\!]_{\mathbf{d}}(T, i) &:= -[\![\phi]\!]_{\mathbf{d}}(T, i) \\
[\![\phi_1 \vee \phi_2]\!]_{\mathbf{d}}(T, i) &:= \max\left([\![\phi_1]\!]_{\mathbf{d}}(T, i), [\![\phi_2]\!]_{\mathbf{d}}(T, i)\right) \\
[\![\bigcirc\phi]\!]_{\mathbf{d}}(T, i) &:= \begin{cases} [\![\phi]\!]_{\mathbf{d}}(T, i+1) & \text{if } i+1 \in N \\ -\infty & \text{otherwise} \end{cases} \\
[\![\phi_1 U_{\mathcal{I}} \phi_2]\!]_{\mathbf{d}}(T, i) &:= \max_{j \text{ s.t. } (t_j - t_i) \in \mathcal{I}} \Bigg(\min\Big([\![\phi_2]\!]_{\mathbf{d}}(T, j), \\
&\qquad \min_{i \leq k < j} [\![\phi_1]\!]_{\mathbf{d}}(T, k) \Big) \Bigg)
\end{aligned}
$$

A trace T satisfies an STL formula ϕ (denoted by $T \models \phi$), if $[\![\phi]\!]_{\mathbf{d}}(T, 0) > 0$. On the other hand, a trace T' falsifies ϕ (denoted by $T' \not\models \phi$), if $[\![\phi]\!]_{\mathbf{d}}(T', 0) < 0$. An overview of the algorithms that can be used to compute $[\![\varphi]\!]_{\mathbf{d}}$ is provided in [41].

Example 3.1:

In order to visualize the specification robustness for an example relevant to this paper, we trained an NN to predict the future position of other vehicles approaching an intersection, similar to the work presented in [15]. Figure 2 shows the scenario under consideration. The ego vehicle is on the x_1 axis and it uses the trained NN in a collision avoidance controller. This simple NN predictor uses the positions of the vehicles as well as the velocity of the adversarial agent (x_2 axis). We consider the requirement that the two vehicles should not be in the intersection at the same time: $\varphi = \Box(\neg(-0.1 \leq x_1 \leq 0.1) \wedge \neg(-0.1 \leq x_2 \leq 0.1))$. In Fig. 3, we present the robustness landscape over the set of initial positions for the two vehicles. It can be observed that the NN controller does a good job avoiding the collisions since we do not have any negative robustness values. Our goal is to use the robustness function to find system behaviors that violate our requirements as discussed in the next section.

B. Robustness-Guided Model Checking (RGMC)

The goal of a *model checking* algorithm is to ensure that all traces satisfy the requirement. The robustness metric can be viewed as a fitness function that indicates the degree to which individual executions of the system satisfy the requirement φ, with positive values indicating that the execution satisfies φ. Therefore, for a given system M and a given requirement φ, the model checking problem is to ensure that for all $T \in \mathcal{L}(M)$, $[\![\varphi]\!]_{\mathbf{d}}(T) > 0$.

Let φ be a given STL property that the system is expected to satisfy. The robustness metric $[\![\varphi]\!]_{\mathbf{d}}$ maps each simulation trace T to a real number r. Ideally, for the STL verification problem, we would like to prove that $\inf_{y \in \mathcal{L}(\Sigma)} \mathcal{R}_{\varphi}(y) > \varepsilon > 0$ where ε is a desired robustness threshold.

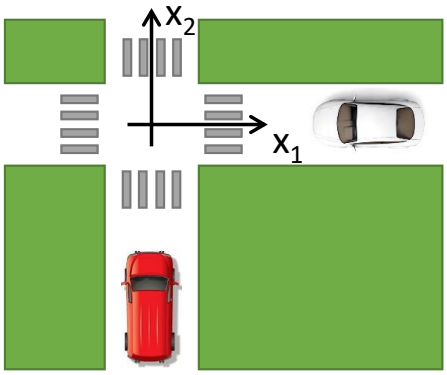


Fig. 2: Simple intersection collision avoidance: an NN has been trained to predict the future position of the vehicles and it is used in the loop for a simple braking controller for the ego vehicle.

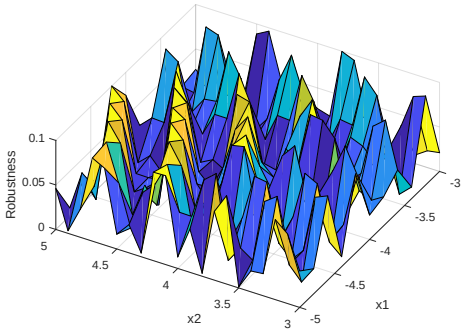


Fig. 3: The resulting robustness landscape for the specification in Example 3.1.

C. Falsification and Critical System Behaviors

In this work, we focus on the task of identifying critical system behaviors, including falsifying traces. To identify falsifying system behaviors, we leverage existing work on *falsification*, which is the process of identifying system traces T that do not satisfy φ. For the STL falsification problem, falsification attempts to solve the problem: Find $T \in \mathcal{L}(\Sigma)$ s.t. $[\![\varphi]\!]_{\mathbf{d}}(T) < 0$. This is done using best effort solutions to the following optimization problem:

$$T^{\star} = \arg \min_{T \in \mathcal{L}(\Sigma)} [\![\varphi]\!]_{\mathbf{d}}(T). \tag{1}$$

If $[\![\varphi]\!]_{\mathbf{d}}(T^{\star}) < 0$, then a counterexample (adversarial sample) has been identified, which can be used for debugging or for training. In order to solve this non-linear non-convex optimization problem, a number of stochastic search optimization methods can be applied (e.g., [45] – for an overview see [46], [47]). We leverage existing falsification methods to identify falsifying examples the autonomous driving system.

D. Covering Arrays

In software systems, there can often be a large number of discrete input parameters that affect the execution path of a program and its outputs. The possible combinations of input values can grow exponentially with the number of parameters. Hence, exhaustive testing on the input space becomes impractical for fairly large systems. A fault in such a system with k parameters may be caused by a specific combination of t parameters, where $1 \leq t \leq k$. One best-effort approach to testing is to make sure that all combinations of any t-sized subset (i.e., all t-way combinations) of the inputs are tested.

A *covering array* is a minimal number of test cases such that any t-way combination of test parameters exist in the list [48]. Covering arrays are generated using optimization-based algorithms with the goal of minimizing the number of test cases. We denote a t-way covering array on k parameters by $CA(t, k, (v_1, ..., v_k))$, where v_i is the number of possible values for the i^{th} parameter. The size of the covering array increases with increasing t, and it becomes an exhaustive list of all combinations when $t = k$. Here, t is considered as the *strength* of the covering array. In practice, t can be chosen such that the generated tests fit into the testing budget. Empirical studies on real-world examples show that more than 90 percent of the software failures can be found by testing 2 to 4-way combinations of inputs [49].

Despite the t-way combinatorial coverage guaranteed by covering arrays, a fault in the system possibly may arise as a result of a combination of a number of parameters larger than t. Hence, covering arrays are typically used to supplement additional testing techniques, like uniform random testing. We consider that because of the nature of the training data or the network structure, NN-based object detection algorithms may be sensitive to a certain combination of properties of the objects in the scene. Figure 4 shows outputs of a DNN-based object detection and classification algorithm for 4 different combinations of vehicle type, vehicle color and pedestrian pants color while all other parameters like position and orientation of the objects are the same. In a comparison between configurations (a) and (b), the vehicle type does not change but the vehicle and pedestrian pants colors change from blue to white. While both the car and the pedestrian are detected in configuration (a), the pedestrian is detected but the car is not detected in configuration (b); however, in a comparison between configurations (b) and (d), if we fix the vehicle and pedestrian pants colors to be white but change the vehicle type, then the car is detected but the pedestrian is not detected. We can also see that the size of the detection box is different between configurations (c) and (d), for which the vehicle type is the same but the vehicle and pedestrian pants colors are different. Our observation is that the characterization of the errors is generally not as simple as saying that all white colored cars are not detected. Instead, the errors arise from some combination of subsets of discrete parameters. Because of this combinatorial aspect of the problem, covering arrays may be a good fit to test DNN-based object detection and classification algorithms. In Sec. V, we describe how Sim-ATAV combines covering arrays to explore discrete and discretized parameters with falsification on continuous parameters.

IV. Requirements

In this section, we provide five STL requirements intended for the autonomous driving system. Each requirement is used to target specific aspects of safety and performance. Also, we describe how analysis results related to each of the requirements can be used to enhance either the controller design or testing phases of the development process.

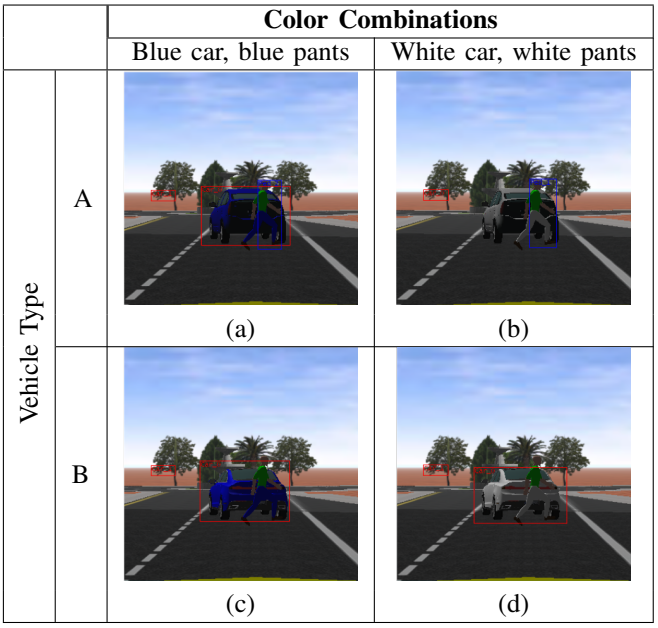


Fig. 4: Specific configurations impacting DNN performance.

A. STL Requirements

The following describes each of the requirements that we use in the sequel to evaluate the autonomous driving system design with our virtual framework. We provide these requirements to illustrate how STL can be used to describe four different types of behavior expectations for an automated driving system: *system-level* safety, *subsystem-level* performance, *subsystem-to-system* safety, and system-level performance (driving comfort) requirements.

Requirement $R1$*:* Vehicle should not collide with an object.

This requirement is an example of a *system-level safety requirement*. It is used to ensure that the ego vehicle does not collide with any object in the environment. Behaviors that do not satisfy this requirement correspond to unsafe performance from the autonomous vehicle. These cases are valuable to identify in simulation, as they can be communicated back to the control designers so that the control algorithms can be improved.

The following provides the STL requirement.

$$R1_i = \Box(\neg \pi_{i,coll})$$

where

$$\pi_{i,coll} = dist(i, ego) < \epsilon_{dist}$$

In the above specification, i corresponds to an object in the environment, such as an agent vehicle or a pedestrian. $dist(i, ego)$ gives the minimum Euclidean distance between the boundaries of the Ego vehicle and the boundaries of object i. The specification basically indicates that the Ego vehicle should not collide with object i.

In practice, we consider a unique requirement for each object in the environment. When the object we are considering is clear from the context, we drop the index i and refer to the requirement $R1$.

Requirement $R2$*:* Sensor should detect visible obstacles.

This requirement is an example of a *subsystem-level requirement*; this particular example can be considered as a requirement on the sensor or perception subsystems. The requirement indicates that the perception system or a specific sensor should not fail to detect an object for an excessive amount of time.

The requirement is as follows.

$$R2_{i,s} = \Box\big((W(i,s) \wedge \neg D(i,s)) \implies \Diamond_{[0,t1]}(D(i,s) \vee \neg W(i,s))\big)$$

Here, we use $W(i,s)$ to mean that object i is physically *visible* to sensor s. For our framework, $s \in \{CCD, lidar, radar, combined\}$, where $combined$ represents the total perception system, which is a fusion of available sensors. Function $D(i,s)$ evaluates to true when sensor s detects object i.

A description of this requirement in natural language could be "it is always true that for any time when object i is visible and not detected by sensor s, there exists an instant, within 0 to $t1$ seconds, that object i is either detected or invisible to the sensor".

When the object i and sensor s are clear from the context, we drop the indices and refer to the requirement $R2$.

Requirement $R3$*:* Localization error should not be too high for too long.

This requirement is another sensor-level requirement and specifies that the localization of an object that is based on a particular sensor should provide sufficient accuracy, within an adequate time after the object becomes visible to the sensor.

The following is the requirement.

$$R3_{i,s} = \Box\big((W(i,s) \wedge (\neg D(i,s) \vee E(i,s) > \epsilon_{err})) \implies \Diamond_{[0,t1]}(\neg W(i,s) \vee (D(i,s) \wedge E(i,s) < \epsilon_{err}))\big)$$

In the requirement, $E(i,s)$ is the difference between object i's location and its location as estimated using information from sensor s. Constant ϵ_{err} is a threshold on the acceptable amount of error between the actual position of i and its estimated position.

To understand the requirement, consider the situation where either an object is not detected (i.e., $\neg D(i,s)$) or there is a large error in the localization of the object (i.e., $E(i,s) > \epsilon_{err}$), and call this a case of *"poor detection"* of the object. Then we can read the requirement as follows: "it is always true that whenever object i is visible to sensor s and is poorly detected by sensor s, there exists an instant, within a time period of 0 to $t1$ seconds, that either object i is invisible to sensor s or the object is detected and the localization error is small, as computed using information from sensor s".

This requirement basically limits the amount of time the sensor error can be greater than a given threshold. When the object i and sensor s are clear from the context, we drop the indices and refer to the requirement $R3$.

Requirement $R4$*:* A sensor-related fault should not lead to a system-level fault.

This is an example of a *subsystem-to-system requirement*. This requirement relates sensor-level behaviors to system-level

behaviors. The purpose is to isolate behaviors where a sensor fault results in a collision. The expectation is that the system as a whole should be robust to failure of a single sensor.

The requirement follows.

$$R4_{i,s} = \Box\neg\Big(\Box_{[0,t1]}\big(\neg\pi_{i,coll} \wedge W(i,s) \wedge (\neg D(i,s) \vee E(i,s) > \epsilon_{err})\big) \wedge \Diamond_{(t1,t2]}\pi_{i,coll}\Big)$$

The above requirement designates that there should not be a period of $t1$ seconds where a visible object is not accurately detected and no collision occurs, followed immediately by a period of length $t2 - t1$ seconds that contains a collision. In other words, the requirement indicates that a system level fault (collision) should not occur within a short time after a sensor fault. A behavior that violates this requirement does not necessarily indicate that the sensor fault *caused* the system fault, but it suggests a correlation, as it points to a behavior wherein the system fault occurs a short time after the sensor fault. Providing behavior examples that violate this requirement can help to pinpoint the cause of system-level faults.

When the object i and sensor s are clear from the context, we drop the indices and refer to the requirement $R4$.

Requirement $R5$: The vehicle should not do excessive braking unnecessarily or too often.

This is a *system-level performance (driving comfort)* requirement, in that it requires that the system not brake unnecessarily or too often, thereby causing discomfort for the passengers.

The requirement follows.

$$R5 = \Box\Big(\neg\Box_{[0,t1]}(B \wedge \neg C) \wedge \neg\big(edge \wedge \Diamond_{(0,t2]}(edge \wedge \Diamond_{(0,t2]} edge)\big)\Big),$$

where

$$edge = B \wedge \bigcirc\neg B.$$

Here, C is a variable that is true when the Ego vehicle is estimated to collide with another object in the environment, based on a simplified model of future behaviors. The simplified model that we use for future trajectory estimation is the Constant Turn Rate and Velocity (CTRV) model [50]. B represents that the amount of braking force applied by the controller exceeds half of the available braking force. The variable $edge$ represents the event of a $true$ value of B followed by a $false$ value in the next time step.

To understand the meaning of requirement $R5$, consider the following part of the requirement:

$$\Box\Big(\neg\Box_{[0,t1]}(B \wedge \neg C)\Big),$$

which requires that the system not apply excessive braking for more than a specific amount of time $(t1)$ while there is no collision predicted. This essentially stipulates that the system should not unnecessarily brake for a prolonged amount of time. Next, consider the second part of requirement $R5$:

$$\Box\Big(\neg\big(edge \wedge \Diamond_{(0,t2]}(edge \wedge \Diamond_{(0,t2]} edge)\big)\Big),$$

which indicates that there should not be an "on-off" behavior, followed by another "on-off" behavior, followed by a third "on-off" behavior, with less than $t2$ seconds between each other. This essentially requires that the brakes not be applied and released too often. Thus, this is a riding comfort requirement.

B. Development Process Support

We describe how requirements $R1$ through $R5$ can be used to support both the controller design and testing phases of the development process.

For all of the requirements, any detected violation (falsification) should be linked back to the conditions that caused the violation.

Consider the first scenario's requirement, $R1$, "Vehicle should not collide with an object": if the vehicle does collide with an object, then we would go back and see what conditions caused such an event, for example, whether the vehicle speed trace exhibited an anomaly or whether the vehicle was moving erratically. Testing for collision avoidance is well established in the field of ADAS. Often inflatable and other destructible targets are employed for convenience; for example, see [51] and Figure 5.

Fig. 5: Robotic pedestrian surrogate target with a Toyota autonomous vehicle.

In the requirement "Sensor should detect visible obstacles" we focus on the detection of an obstacle as operational imperative. If the sensor fails to detect within a time interval, then the requirement will be violated. This is essentially the sensor-level requirement (visible but not detected), and test engineers can set a real-world experiment to verify it relatively easy because it is decoupled from others (one-term inequality, sensor by sensor).

The requirement "Localization error should not be too high for too long" is important to verify (falsify) for both ego-location and identification of positions of other agents in the environment. Placing an ego-vehicle in the correct pose on the road is usually not achieved by simply relying on GPS signal processing, due to the GPS tendency to "jump" unpredictably, but instead by estimating and dynamically refining the pose through landmark observations, such as road edges, vertical elements such as light poles, and signs. Assuming that the ego-vehicle localization is done with sufficient accuracy, the remaining task of localizing is to make sure that the location of other agents, especially those in the planned path of the ego

vehicle, are estimated with sufficient accuracy. Often a grid-based representation centered on the ego-vehicle is employed (e.g., [52]). Estimating $E(i, s)$ in $R3$ is not trivial, but practical approaches exist that can be used by test engineers (e.g., [53]).

The requirement "A sensor-related fault should not lead to a system-level fault" is a form of robustness requirement. This is similar to a requirement that the system should have no "single point of failure", which enforces that the failure of any single component will not cause the system to fail (for example, see [54]). We make an important clarification that is practical but limiting in scope: that no failure should occur within the specified (short) time after the fault. Test engineers could readily use examples of behavior provided in the course of falsifying this requirement.

Lastly, the requirement "The vehicle should not brake too often" is an example of a possible set of requirements designed to establish how comfortable the ride in the vehicle is. It is known that autonomous vehicles could induce motion sickness in passengers if the vehicle control system does not comply with human physiology [55], [56]. A better requirement may well be developed using fuzzy set theory and further refined for a specific target group of passengers (e.g., elderly people). An alternative requirement could be defined by counting the number of occurrences of an event within a total time period, instead of relating one occurrence to another. Such a requirement can be defined as a Timed Propositional Temporal Logic (TPTL) specification. TPTL is a variant of temporal logic and it is also supported in our framework [57].

V. Framework

We describe Sim-ATAV, a framework for performing testing and analysis of autonomous driving systems in a virtual environment. The simulation environment used in Sim-ATAV includes a vehicle perception system, a vehicle controller, and a model of the physical environment. The perception system processes data from three sensor systems: CCD camera images, lidar, and radar. The framework uses freely available and low cost tools and can be run on a standard desktop PC. Later, we demonstrate how Sim-ATAV can be used to implement traditional testing approaches, as well as advanced automated testing approaches that were not previously possible using existing frameworks.

Figure 6 shows an overview of the simulation environment. The environment consists of a Simulator and a Vehicle Control system. The Simulator contains models of the ego vehicle, agents, and other objects in the environment (e.g., roads, buildings). The Simulator outputs sensor data to the Vehicle Control system. The sensor data includes representations of CCD camera, lidar, and radar data. Simple models of the sensors are used to produce the sensor data. For example, synthetic CCD camera images are rendered by the Simulation system, as if they came from a camera mounted on the front of the ego vehicle. The Vehicle Control system contains models of the Perception System, which performs sensor data processing and sensor fusion. The Controller uses the output of the Perception System to make decisions about how to actuate the AV system. Actuation commands are sent from the Controller to the Simulator.

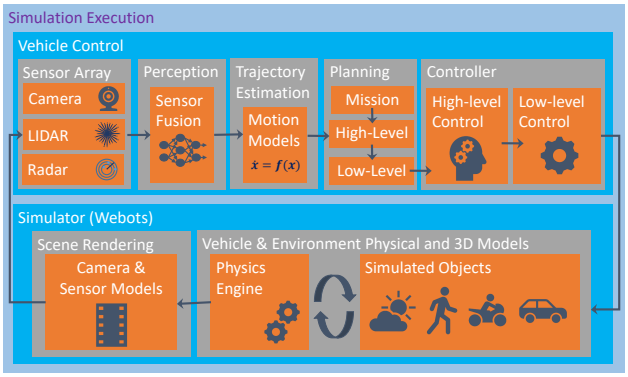


Fig. 6: Overview of the simulation environment.

Simulations proceed iteratively. At each instant, sensor data is processed by the Vehicle Control, which then makes an actuation decision. The actuation decision is then transmitted back to the Simulator, which uses the actuation commands to update the physics for the next time instant. This process is repeated until a designated time limit has been reached.

The Vehicle Control system is implemented in Python. We use simplified algorithms to implement the subsystems of the vehicle control, which is sufficient in this case, as the purpose of this investigation is to evaluate new *testing* methodologies and not to evaluate a real AV control design; however, we note that it is straightforward to replace our algorithms with production versions to test real control designs.

To process CCD image data, we use a lightweight DNN, SqueezeDet, which performs object detection and classification [58]. SqueezeDet is implemented in TensorFlow™[59], and it outputs a list of object detection boxes with corresponding class probabilities. This network was originally trained on real image data from the KITTI dataset [11] to achieve accuracy comparable to the popular AlexNet classifier [16]. We further train this network on the virtual images generated in our framework. Fig. 7 shows an example output from SqueezeDet, based on a synthetic image produced by our simulator. The image shows two vehicles correctly detected and classified, along with a portion of a shadow that is incorrectly classified as a vehicle.

To process lidar point cloud data, we cluster the received points based on their positions and estimate existence and types of the objects based on the dimensions of the clusters. We implement a simple sensor fusion algorithm that relates and merges the object detections from camera, lidar, and radar with a simple logic. It also utilizes the expected current positions of previously detected objects. The object states estimated by the sensor fusion algorithm are used to estimate the future trajectories of the objects using the CTRV model [50].

Fig. 8 illustrates outputs from the sensor fusion system. In the figure, the solid yellow box in the middle represents the Ego vehicle. Yellow circles in front of the ego vehicle represent the estimated future trajectory of the Ego vehicle. Small white dots represent lidar point cloud data. The colored dots and rectangles represent detected objects, with their estimated orientation indicated with a white line in front of them. Expected future positions of agent vehicles with respect

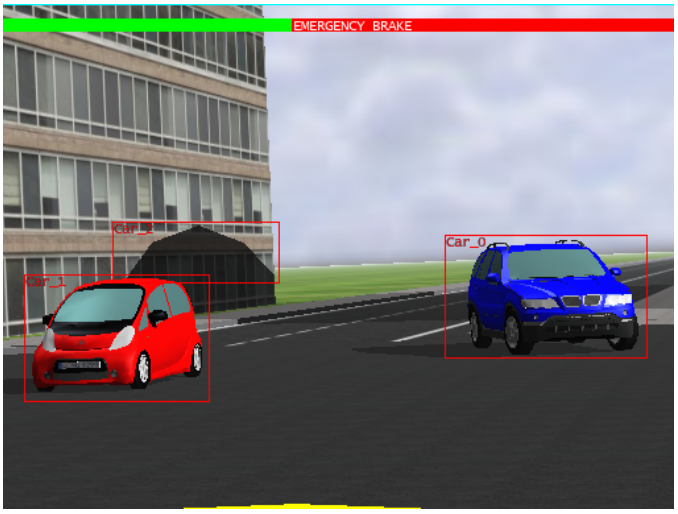


Fig. 7: Outputs from the SqueezeDet DNN, based on a synthesized camera image.

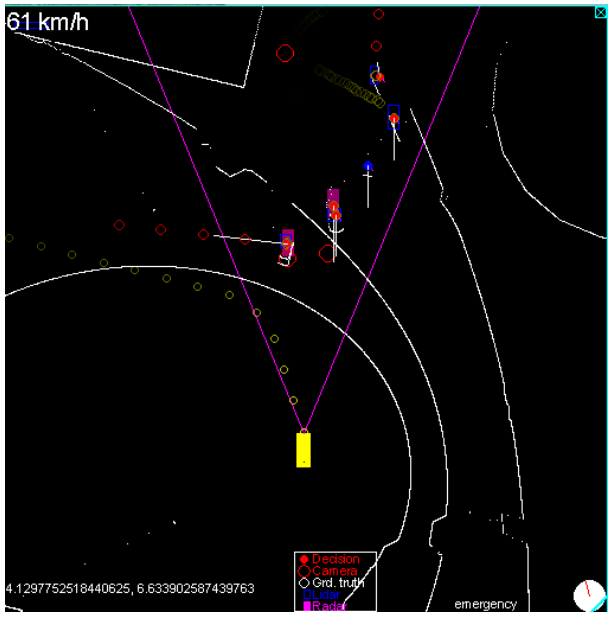


Fig. 8: Sensor fusion outputs.

to the ego vehicle are represented by red circles.

Our simple planner takes the high level target path and target speed and the outputs of sensor fusion and trajectory estimation algorithms. It assigns collision risk levels to the target objects with a simple logic and outputs the risk assessments and a target speed, which depends on the target speed of the mission or other factors, such as the distance to a sharp turn ahead.

Our control algorithm implements simple path and speed tracking and *collision avoidance* features. The controller receives the outputs of the planner. When there is no collision risk, the controller drives the car with the target speed and on the target path. When a future collision with an object is predicted, it applies the brakes at a level commensurate with the risk assigned to the object.

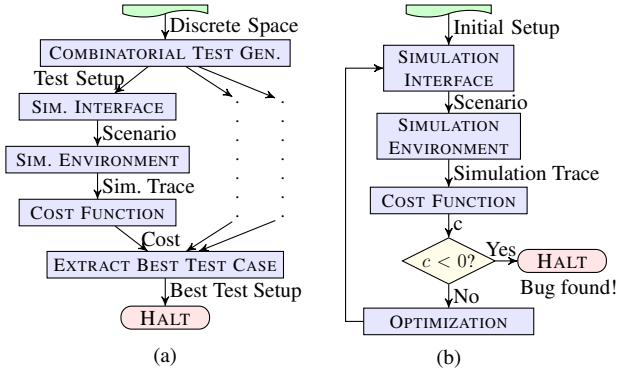


Fig. 9: Flowcharts illustrating the combinatorial testing (a) and falsification (b) approaches.

The environment modeling framework is implemented in Webots [60], a robotic simulation framework that models the physical behavior of robotic components, such as manipulators and wheeled robots, and can be configured to model autonomous driving scenarios. In addition to modeling the physics, a graphics engine is used to produce images of the scenarios. In Sim-ATAV, the images rendered by Webots are configured to correspond to the image data captured from a virtual camera that is attached to the front of a vehicle.

The process used by Sim-ATAV for test generation and execution for discrete and discretized continuous parameters is illustrated by the flowchart shown in Fig. 9-(a). Sim-ATAV first generates test cases that correspond to scenarios defined in the simulation environment using covering arrays as a combinatorial test generation approach. The scenario setup is communicated to the simulation interface using TCP/IP sockets. After a simulation is executed, the corresponding simulation trace is received via socket communication and evaluated using a cost function. Among all discrete test cases, the most promising one is used as the initial test case for the falsification process shown in Fig. 9-(b). For falsification, the result obtained from the cost function is used in an optimization setting to generate the next scenario to be simulated. For this purpose, we used S-TaLiRo [43], which is a MATLAB® toolbox for falsification of CPSs. Similar tools, such as Breach [44], can also be used in our framework for the same purpose.

VI. Testing Application

In this section, we present an evaluation of our Sim-ATAV framework using three separate driving scenarios. The scenarios are selected to be both challenging for the autonomous driving system and also analogous to plausible driving scenarios experienced in real-world situations. In general, the system designers will need to identify crucial driving scenarios, based on intuition about challenging situations, from the perspective of the autonomous vehicle control system. A thorough simulation-based testing approach will include a wide array of scenarios that exemplify critical driving situations.

For each of the following scenarios, we consider a subset of the requirements presented in Sec. IV and describe how to use the results to enhance the development process. We conclude the section with a summary of the results.

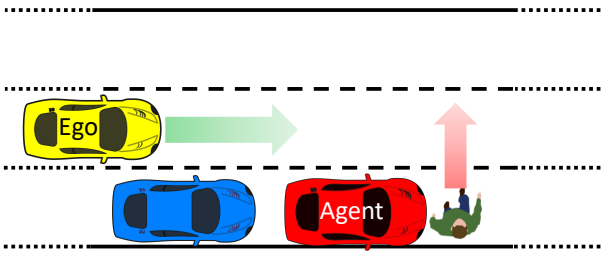


Fig. 10: Overview of the scenario 1.

Scenario 1

The first scene that we consider is a straightaway section of a two-lane road, as illustrated in Fig. 10. Several cars are parked on the right-hand side of the road, and a pedestrian is jay-walking in front of one of the cars, passing in front of the Ego car from right to left. We call this driving scenario model M_1. The scenario simulates a similar setup to the Euro NCAP Vulnerable Road User (VRU) protection test protocols [4].

Several aspects of the driving scenario are parameterized, meaning that their values are fixed for any given simulation by appropriately selecting the model parameters $\mathbf{p}_d$. The parameters we use for this scenario are as follows:

- Initial speed and lateral position of the Ego vehicle inside its lane;
- Walking speed of the pedestrian;
- The model of Agent car 1, which is next to the pedestrian;
- R, G, B values for the colors of Agent car 1;
- R, G, B values for the pedestrian's shirt and pants.

We choose the parameters such that their specific combinations could be challenging to a DNN-based pedestrian detection system that relies on CCD camera images. We also choose some of the parameter ranges so that the scenario is physically challenging for the brake performance.

We evaluate Model M_1 against three of the requirements from Sec. IV: $R1$, $R2$, and $R4$. These include the system-level requirement, the sensor-level requirements, and the sensor-to-system-level requirement. We use this collection of requirements for Model M_1 to demonstrate how we can automatically identify each type of behavior using our framework.

Scenario 2

The next scenario involves a left turn maneuver by the ego vehicle in a controlled intersection, as illustrated in Fig. 11. An agent vehicle *(Agent 1)* in the opposing lane unexpectedly passes through the intersection, against a red light, potentially causing a collision with the Ego vehicle. There is also another agent car *(Agent 2)*, which is making a legal left turn from the opposing lane. It is incumbent on the Ego vehicle to take action to avoid colliding with the agent vehicles. We call the model of this scenario M_2.

For this experiment we choose parameters such that the position of Agent 2, or trajectory followed by Agent 1, in combination with the behavior of the Ego, may result in poor performance from the sensor processing or trajectory estimation systems. The following variables are parameterized for this model:

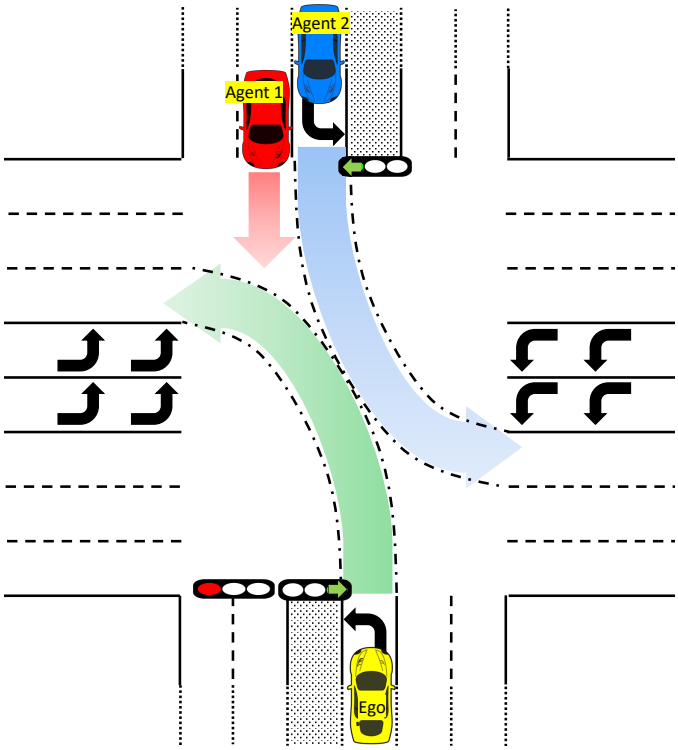


Fig. 11: Overview of the scenario 2.

- Ego vehicle initial speed and initial distance to the intersection;
- Agent 1 initial distance to the intersection, initial target speed, target speed when approaching the intersection, target speed inside the intersection, initial lateral position w.r.t its lane center, target lateral position w.r.t its lane center when approaching the intersection, and target lateral position w.r.t its lane center inside the intersection;
- Agent 2 initial lateral position w.r.t its lane center, speed, and initial distance to the intersection.

We evaluate Model M_2 against requirement $R4$. The idea in using the sensor-to-system-level requirement is that it is relatively easy, in general, to find behaviors that result in a collision for Model M_2, but many collision cases are not interesting for the designers. This could be because, for example, the agent car is moving too quickly for the ego vehicle to avoid. This would be a behavior that is not necessarily caused by any specific incorrect behavior on the part of the ego vehicle. Instead, we use $R4$ to identify behaviors where there is a collision that is directly correlated to unacceptable performance from the sensor processing system; in a sense, these are cases where the sensor data processing or future trajectory estimation system is at fault for the collision. These are more valuable cases, in that they can more easily be used to debug specific aspects of the ego vehicle control algorithms.

Scenario 3

In this last scenario, the ego vehicle is making a left turn through an intersection, while an agent vehicle in the opposing lane is also making a left turn. This scene is similar to the Scenario 2, as depicted in Fig. 11, except that Agent 1 is

not present in this scenario, only Agent 2, which we refer to as the agent vehicle for this scenario. If both ego and agent vehicles are not accurately regulating their trajectories during this maneuver, a collision may occur. We call the model of this scenario M_3.

In this scenario, we search over target trajectories of the ego and agent vehicles. Below are the parameters that we use:

- Ego vehicle initial speed, target lateral position w.r.t its lane center when entering the intersection, distance traveled inside the intersection before starting its left turn, target lateral position w.r.t its lane center when exiting the the intersection, and distance traveled inside the intersection after finishing its left turn;
- Agent vehicle speed, target lateral position w.r.t its lane center when entering the intersection, distance traveled inside the intersection before starting its left turn, target lateral position w.r.t its lane center when exiting the the intersection, and distance traveled inside the intersection after finishing its left turn.

We evaluate Model M_3 against requirement $R5$. The purpose of considering the performance requirement $R5$ in this case, is that scenario M_3 is difficult to falsify. That is, due to the specific parameter ranges selected for the scenario, it is unlikely that the ego vehicle will collide with the agent vehicle. Instead, in this case, we are interested to identify situations where the emergency braking system unnecessarily decelerates the ego vehicle, causing unacceptable performance, from a ride-quality perspective. The scenario can easily lead to unnecessary braking, as the ego and agent vehicles momentarily move toward each other during their left turn maneuvers, which can cause the emergency braking algorithm to decide, incorrectly, that a collision is imminent. This type of case can be useful as feedback to designers, as it can highlight controller behaviors that are too conservative, at the expense of ride quality.

Summary of Test Results

We present results from experiments demonstrating the application of our framework to the scenarios and requirements described above. Table I summarizes the results. Indicated in the table for each case study are the requirements used to test each model, the testing approach used, the set of active sensors used, and a summary of the results. We describe the results in detail below.

Covering array and falsification on Model M_1: In our previous work [17], we proposed and studied the effectiveness of a testing approach that first uses covering arrays to discover critical regions, based on a set of discrete parameters, then uses those results as the initial points for robustness guided falsification. Here, we apply that approach on model M_1 for 3 different requirements, $R1$, $R2$ and $R4$. In model M_1, we focus on the camera sensor and DNN-based object detection and classification algorithm. Because of this, most of our parameters are colors of pedestrian clothing and the agent vehicle, as described in Sec. IV. We first execute 195 covering array tests and collect simulation trajectories. Then, we compute the *robustness* values for those trajectories, with respect to the requirements $R1$, $R2$ and $R4$. Finally, for each requirement, starting from the case with the smallest positive robustness value, we try to find as many additional falsifications as possible, within a maximum of 300 extra simulations, by using a falsification approach that uses simulated annealing to perform the optimization.

For requirement $R1$, 67 cases were falsified from the covering array tests (i.e., 67 of the 195 cases did not satisfy $R1$). Starting from 7 of the remaining (non-falsifying) cases from the covering array tests, 5 additional falsifying cases were discovered using falsification. For requirement $R2$, 65 cases were falsified from the covering array cases, with an additional 8 cases discovered during the falsification step. For requirement $R4$, 67 cases were falsified during the covering array step, with 12 more cases discovered during the falsification step.

These results demonstrate that we can automatically identify test cases that violate specific sensor-level, system-level, and sensor-to-system level requirements. These test cases can be fed back to the designers to improve the perception or control design or can be used as guidance to identify challenging scenarios to be used during the testing phase.

Analysis of robustness values on the falsification of Model M_2: The *robustness* value, which is described in Sec. III, for a trajectory with respect to the requirement is automatically computed in Sim-ATAV. This computation is performed by the S-TaLiRo tool [42] and is used to guide the test cases towards a falsification.

We use the results of falsification on Model M_2 to show, in Fig. 12, how the robustness value changes over time and finally becomes negative, which indicates falsification of the requirement. In this case, Sim-ATAV was able to find a falsifying example in 58 simulations. Because the cost function gradients are not computable, we use a stochastic global optimization technique, Simulated Annealing (SA). The blue line shows the robustness value for each simulation. We can observe that the robustness value per simulation run is not monotonic. This is due to the stochastic nature of the optimizer; however, the achieved minimum robustness up to the current simulation is a non-increasing function, which shows the best robustness achieved after each simulation. As soon as the framework finds a test case that causes a negative robustness value, it stops the search and reports the falsifying example.

Fig. 13 shows images from the simulation execution of a falsifying example for model M_2 with respect to the requirement $R2$. Between the time corresponding to Fig. 13-(a) to Fig. 13-(b), the red car approaching from the opposite side is driving on a path such that there will be a future collision with the Ego vehicle; however, due to incorrect locatization of the agent vehicle, the Ego vehicle is not able to correctly predict the future trajectory of the agent vehicle, and so it does not predict a collision. Hence it continues without taking action to avoid the collision. Starting from the moment shown on Fig. 13-(c), the Ego vehicle predicts the collision and starts applying emergency braking; however, because it takes action too late, the Ego vehicle cannot avoid a collision with the agent vehicle, as shown in Fig. 13-(d).

We note that, even for cases that are non-falsifying, the

	Model M_1			Model M_2	Model M_3
Requirement	$R1$	$R2$	$R4$	$R4$	$R5$
Testing Modality	CA+ Falsification	CA+ Falsification	CA+ Falsification	Falsification	Falsification
Active Sensors	CCD	CCD	CCD	CCD, Radar, LIDAR	CCD, LIDAR
Computation Time	CA: 2h, 10min.			2h, 3min.	9h, 40min.
	Fals.:3h, 33min.	3h, 35min.	3h 34min.		
No. Simulations	CA: 195			58	232
	Fals.:300	300	300		
Falsification Obtained	67 by CA + 5 by falsification	65 by CA + 8 by falsification	67 by CA + 12 by falsification	✓	✓
Application of Results	Lowest robustness cases used to create critical tests.			Falsifying cases relate to processing of specific sensor; aids in controller design improvement.	Poor performance cases used to improve controller design in modeling phase.

TABLE I: Results from autonomous driving tests using virtual framework.

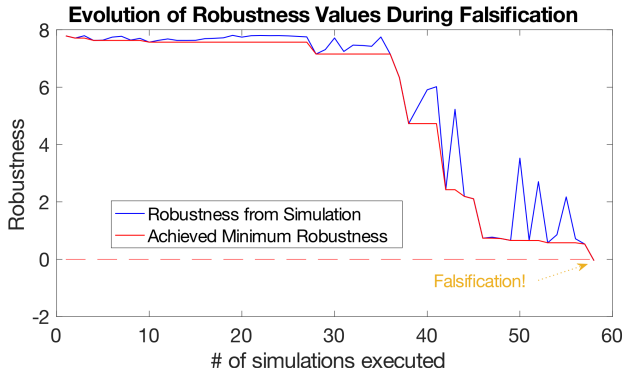


Fig. 12: Robustness guided falsification utilizes global optimization techniques to guide the test cases toward falsification.

robustness values are useful for the system designer's analysis, as behaviors with low robustness value are "close to" violating the requirement and therefore correspond to cases that may require closer attention.

A visual analysis of a falsifying simulation trajectory from Model M_3: As presented in Table I, Sim-ATAV was able to find a falsifying example for model M_3 with respect to the STL requirement $R5$ in 232 simulations. We present a visual analysis of the falsifying test result. Note that this analysis is done automatically in the framework, and corresponding satisfaction/falsification of the requirement is returned to the user, along with the *robustness* value that shows the signed distance to the boundary of satisfaction or falsification. The type of visual analysis we present here may be useful for the system designers to understand the reason behind the falsification (or satisfaction) of a requirement, which can be helpful for debugging or improving the design. For this analysis, we use the definitions and notation introduced in Section IV.

Fig. 14 shows a part of the simulation trajectory of Model M_3 for a time window around the falsification instance, together with the corresponding logic evaluations of the predicates related to the subformulas in Requirement $R5$. In the top plot in Fig. 14, the red solid line is the estimated future minimum distance between the ego vehicle and Agent vehicle 1, with respect to the simulation time. This estimation is based on the ground truth information collected from the simulation and utilizes the CTRV model at each time step of the simulation to compute the expected estimation that is described in $R5$. For this example, we define the variable C that is used in $R5$ as $(d_{f,min} < 0.5)$, where $d_{f,min}$ represents the expected minimum future distance; the dashed horizontal red line in Fig. 14 located at $0.5m$ is the threshold minimum future distance for a collision estimation. The values of $t1$ and $t2$ are respectively defined as 0.6 and 0.5 in this example. Since $d_{f,min}$ is never less than 0.5 in this case, the collision estimation variable C, which is represented by the black solid line in the top plot, is always false.

The middle plot presents a similar evaluation for computing the variable B used in $R5$, which represents excessive braking. This evaluation uses the collected actual normalized brake power data, say br, from the simulation and computes the logical variable $B = (br > 0.5)$. The solid and dashed red lines represent br and the threshold value 0.5, respectively. The solid black line shows the value of B with respect to time.

The bottom plot in Fig. 14 shows the value of the variable $edge$ that is defined for the requirement $R5$ with respect to the simulation time.

The first part of the requirement $R5$, which was defined as $\square\Big(\neg\square_{[0,t1]}(B \wedge \neg C)\Big)$ in Section IV, would evaluate to false if and only if there would exists a time window of $t1$ seconds such that B is always true and C is always false. Focusing on the values of C and B from the top two plots in Fig. 14, we can see that although C is always false, because there is no time window of $t1 = 0.6s$ in which B is always true, the first part of the requirement evaluates to true. This means this execution of model M_3 satisfies the first part of the requirement $R5$.

The second part of the requirement $R5$, which is defined as $\square\Big(\neg\big(edge \wedge \Diamond_{(0,t2]}(edge \wedge \Diamond_{(0,t2]} edge)\big)\Big)$ evaluates to false if and only if there exists a series of three falling edges of B ($edge$), such that one occurrence of $edge$ follows another within a time window of $t2$ seconds. As we see in the bottom

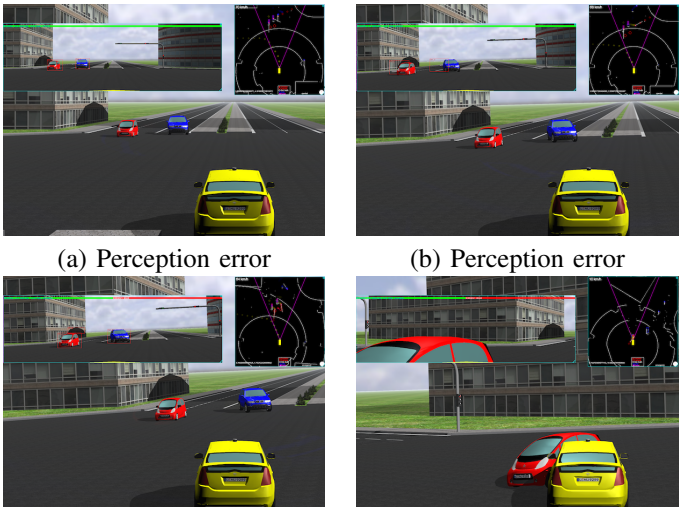
plot of Fig. 14, at time $5.6s$ it is true that there exists an $edge$ and it is also true that there exists another $edge$ within the time window of 0 to $0.5s$ following this moment (occurring at $5.85s$). Hence the inner $(edge \wedge \Diamond_{(0,t2]} edge)$ inside the above formula evaluates to true at time $5.6s$. If we call this event $e1$, the overall formula will evaluate to false if there exists an $edge$ that is followed by event $e1$ in a time window between 0 and $t2 = 0.5s$. This happens at time $5.46s$, which is the moment that there exists an $edge$ followed by event $e1$ at $0.14 \in (0, 0.5]$ seconds, *where the event* $e1$ *is defined as an* $edge$ *followed by another* $edge$ *within* $t \in (0, 0.5]$ *seconds*. Hence, the second part of the requirement $R5$ evaluates to false, and as a result, $R5$ evaluates to false at time $5.46s$, since it is a conjunction of parts 1 and 2. In other words, the system falsifies (does not satisfy) the requirement $R5$.

VII. Conclusions

We demonstrated a simulation-based adversarial test generation framework for autonomous vehicles. The framework works in a closed-loop fashion, where the system evolves in time with the feedback cycles from the autonomous vehicle's controller. The framework includes models of lidar and radar sensor behaviors, as well as a model of the CCD camera sensor inputs. CCD camera images are rendered synthetically by our framework and processed using a pre-trained deep neural network (DNN). Using our framework, we demonstrated a new effective way of finding a critical vehicle behaviors by using 1) covering arrays to test combinations of discrete parameters and 2) simulated annealing to find corner-cases.

Future work will include using identified counterexamples to retrain and improve the DNN-based perception system. Additionally, the scene rendering will be made more realistic by using other scene rendering tools, such as those based on state-of-the-art game engines.

Also, we note that the formal requirements that we considered were provided as an example of the type used when



(a) Perception error (b) Perception error

(c) Emergency braking (d) Collision

Fig. 13: Time-ordered images from the falsifying example on model M_1.

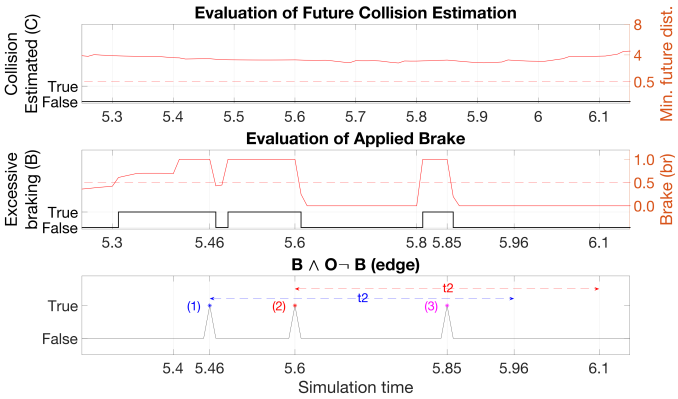


Fig. 14: Analysis of falsification for Model M_3.

employing a requirements-driven development approach based on a temporal logic language, which is a formalism that may be unfamiliar to many test engineers. Future research will include investigating ways to automatically produce formal requirements based on requirements given in more traditional forms.

References

[1] "NHTSA Federal Automated Vehicles Policy," https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016, accessed: 2018-09-11.

[2] SAE, "Guidelines for safe on-road testing of sae level 3, 4, and 5 prototype automated driving systems (ADS), document j3018_201503," https://www.sae.org/standards/content/j3018_201503/, accessed: 2018-09-12.

[3] A. Christensen, A. Cunningham, J. Engelman, C. Green, C. Kawashima, S. Kiger, D. Prokhorov, L. Tellis, B. Wendling, and F. Barickman, "Key considerations in the development of driving automation systems," in *24th Enhanced Safety of Vehicles Conferences*, 2015.

[4] "Euro NCAP," https://www.euroncap.com/en, accessed: 2018-09-11.

[5] "Pegasus," https://www.pegasusprojekt.de/en/home, accessed: 2018-09-11.

[6] J. E. Stellet, M. R. Zofka, J. Schumacher, T. Schamm, F. Niewels, and J. M. Zöllner, "Testing of advanced driver assistance towards automated driving: A survey and taxonomy on existing approaches and open questions," *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 1455–1462, 2015.

[7] M. R. Zofka, M. Essinger, T. Fleck, R. Kohlhaas, and J. M. Zöllner, "The sleepwalker framework: Verification and validation of autonomous vehicles by mixed reality lidar stimulation," in *SIMPAR*. IEEE, 2018, pp. 151–157.

[8] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Trans. Safety*, vol. 4, pp. 15–24, 04 2016. [Online]. Available: https://doi.org/10.4271/2016-01-0128

[9] W. Burgard, U. Franke, M. Enzweiler, and M. Trivedi, "The Mobile Revolution - Machine Intelligence for Autonomous Vehicles (Dagstuhl Seminar 15462)," *Dagstuhl Reports*, vol. 5, no. 11, pp. 62–70, 2016. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2016/5764

[10] W. Wachenfeld, P. Junietz, R. Wenzel, and H. Winner, "The worst-time-to-collision metric for situation identification," in *2016 IEEE Intelligent Vehicles Symposium, IV 2016, Gotenburg, Sweden, June 19-22, 2016*, 2016, pp. 729–734.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[12] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Advances in neural information processing systems*, 1989, pp. 305–313.

[13] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.

[14] L. Chi and Y. Mu, "Deep steering: Learning end-to-end driving model from spatial and temporal visual cues," *arXiv preprint arXiv: 1708.03798*, 2017.

[15] M. Strickland, G. Fainekos, and H. B. Amor, "Deep predictive models for collision risk assessment in autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] C. E. Tuncali, G. Fainekos, H. Ito, and J. Kapinski, "Simulation-based Adversarial Test Generation for Autonomous Vehicles with Machine Learning Components," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[18] W. G. Najm, R. Ranganathan, G. Srinivasan, J. S. Toma, E. Swanson, and A. B. D. Smith, "Description of light-vehicle pre-crash scenarios for safety applications based on vehicle-to-vehicle communications," DOT HS 811 731, Tech. Rep., 2013.

[19] D. Zhao, Y. Guo, and Y. J. Jia, "TrafficNet: An open naturalistic driving scenario library," in *20th IEEE International Conference on Intelligent Transportation Systems, ITSC*, 2017.

[20] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 595–607, 2017.

[21] S. M. Loos, A. Platzer, and L. Nistor, "Adaptive cruise control: Hybrid, distributed, and now formally verified," in *Formal Methods*, ser. LNCS, vol. 6664. Springer, 2011, pp. 42–56.

[22] M. Althoff, D. Althoff, D. Wollherr, and M. Buss, "Safety verification of autonomous vehicles for coordinated evasive maneuvers," in *IEEE Intelligent Vehicles Symposium*, 2010.

[23] C. E. Tuncali, T. P. Pavlic, and G. Fainekos, "Utilizing S-TaLiRo as an automatic test generation framework for autonomous vehicles," in *IEEE Intelligent Transportation Systems Conference*, 2016.

[24] M. Althoff and S. Lutz, "Automatic generation of safety-critical test scenarios for collision avoidance of road vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[25] C. E. Tuncali, S. Yaghoubi, T. P. Pavlic, and G. Fainekos, "Functional gradient descent optimization for automatic test case generation for vehicle controllers," in *IEEE International Conference on Automation Science and Engineering*, 2017.

[26] M. O'Kelly, H. Abbas, and R. Mangharam, "Computer-aided design for safe autonomous vehicles," in *2017 Resilience Week (RWS)*, 2017.

[27] B. Kim, Y. Kashiba, S. Dai, and S. Shiraishi, "Testing autonomous vehicle software in the virtual prototyping environment," *Embedded Systems Letters*, vol. 9, no. 1, pp. 5–8, 2017.

[28] B. Kim, A. Jarandikar, J. Shum, S. Shiraishi, and M. Yamaura, "The SMT-based automatic road network generation in vehicle simulation environment," in *International Conference on Embedded Software (EMSOFT)*. ACM, 2016, pp. 18:1–18:10.

[29] C. Fan, B. Qi, and S. Mitra, "Data-driven formal reasoning and their applications in safety analysis of vehicle autonomy features," *IEEE Design Test*, vol. 35, no. 3, pp. 31–38, June 2018.

[30] M. O'Kelly, H. Abbas, S. Gao, S. Shiraishi, S. Kato, and R. Mangharam, "APEX: Autonomous vehicle plan verification and execution," in *SAE World Congress*, 2016.

[31] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, 2011, pp. 1918–1921.

[32] V. John, K. Yoneda, B. Qi, Z. Liu, and S. Mita, "Traffic light recognition in varying illumination using deep learning and saliency map," in *Proceedings of the 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, 2014.

[33] A. Angelova, A. Krizhevsky, and V. Vanhoucke, "Pedestrian detection with a large-field-of-view deep network," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '15)*, 2015.

[34] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *40th International Conference on Software Engineering (ICSE)*, 2018.

[35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ACM Asia Conference on Computer and Communications Security*. ACM, 2017.

[36] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *arXiv:1701.00160v3*, 2017.

[37] S. Dutta, S. Jha, S. Sankaranarayanan, and A. Tiwari, "Learning and verification of feedback control systems using feedforward neural networks," in *Analysis and Design of Hybrid Systems*, 2018.

[38] T. Dreossi, A. Donze, and S. A. Seshia, "Compositional falsication of cyber-physical systems with machine learning components," in *NASA Formal Methods (NFM)*, ser. LNCS, vol. 10227. Springer, 2017, pp. 357–372.

[39] T. Dreossi, S. Jha, and S. A. Seshia, "Semantic adversarial deep learning," *arXiv:1804.07045v2*, 2018.

[40] H. Abbas, M. O'Kelly, A. Rodionova, and R. Mangharam, "Safe at any speed: A simulation-based test harness for autonomous vehicles," in 7^{th} *International Workshop on Cyber-Physical Systems (CyPhy)*, 2017.

[41] E. Bartocci, J. Deshmukh, A. Donzé, G. Fainekos, O. Maler, D. Nickovic, and S. Sankaranarayanan, "Specification-based monitoring of cyber-physical systems: A survey on theory, tools and applications," in *Lectures on Runtime Verification - Introductory and Advanced Topics*, ser. LNCS. Springer, 2018, vol. 10457, pp. 128–168.

[42] G. E. Fainekos and G. J. Pappas, "Robustness of temporal logic specifications," in *Formal Approaches to Testing and Runtime Verification*, ser. LNCS, vol. 4262. Springer, 2006, pp. 178–192.

[43] G. Fainekos, S. Sankaranarayanan, K. Ueda, and H. Yazarel, "Verification of automotive control applications using s-taliro," in *Proceedings of the American Control Conference*, 2012.

[44] A. Donze and O. Maler, "Robust satisfaction of temporal logic over real-valued signals," in *Formal Modelling and Analysis of Timed Systems*, ser. LNCS, vol. 6246. Springer, 2010.

[45] H. Abbas, G. E. Fainekos, S. Sankaranarayanan, F. Ivancic, and A. Gupta, "Probabilistic temporal logic falsification of cyber-physical systems," *ACM Transactions on Embedded Computing Systems*, vol. 12, no. s2, May 2013.

[46] B. Hoxha, H. Bach, H. Abbas, A. Dokhanchi, Y. Kobayashi, and G. Fainekos, "Towards formal specification visualization for testing and monitoring of cyber-physical systems," in *International Workshop on Design and Implementation of Formal Tools and Systems*, 2014.

[47] J. Kapinski, J. V. Deshmukh, X. Jin, H. Ito, and K. Butts, "Simulation-based approaches for verification of embedded control systems: An overview of traditional and advanced modeling, testing, and verification techniques," *IEEE Control Systems Magazine*, vol. 36, no. 6, pp. 45–64, 2016.

[48] A. Hartman, "Software and hardware testing using combinatorial covering suites," *Graph theory, combinatorics and algorithms*, vol. 34, pp. 237–266, 2005.

[49] D. R. Kuhn, R. N. Kacker, and Y. Lei, *Introduction to combinatorial testing*. CRC press, 2013.

[50] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *2008 11th International Conference on Information Fusion*, June 2008, pp. 1–6.

[51] D. J. LeBlanc, M. Gilbert, S. Stachowski, D. Blower, C. A. C. Flannagan, S. Karamihas, and W. T. B. andRini Sherony, "Advanced surrogate target development for evaluating pre-collision systems," in *23rd Enhanced Safety of Vehicles Conferences*, 2013.

[52] A. Petrovskaya and S. Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Autonomous Robots*, vol. 26, no. 2, pp. 123–139, Apr 2009.

[53] B. Grabe, T. Ike, and M. Hötter, "Evidence based evaluation method for grid-based environmental representation," in *FUSION*. IEEE, 2009, pp. 1234–1240.

[54] "ISO Functional Safety," https://en.wikipedia.org/wiki/ISO_26262, accessed: 2018-09-11.

[55] M. Elbanhawi, M. Simic, and R. Jazar, "In the passenger seat: Investigating ride comfort measures in autonomous cars," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 3, pp. 4–17, Fall 2015.

[56] P. Green, "Motion sickness and concerns for self-driving vehicles: A literature review," http://umich.edu/ driving/publications/Motion-Sickness–Report-061616pg-sent.pdf, accessed: 2018-09-11.

[57] A. Dokhanchi, B. Hoxha, C. E. Tuncali, and G. Fainekos, "An efficient algorithm for monitoring practical tptl specifications," in *Formal Methods and Models for System Design (MEMOCODE), 2016 ACM/IEEE International Conference on*. IEEE, 2016, pp. 184–193.

[58] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," *arXiv preprint arXiv:1612.01051*, 2016.

[59] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint:1603.04467*, 2016.

[60] O. Michel, "Cyberbotics ltd. Webots: professional mobile robot simulation," *International Journal of Advanced Robotic Systems*, vol. 1, no. 1, p. 5, 2004.