Journal of Systems Architecture 110 (2020) 101766

Contents lists available at ScienceDirect

Journal of Systems Architecture

journal homepage: www.elsevier.com/locate/sysarc

Attacking vision-based perception in end-to-end autonomous driving models☆

Adith Boloor[a,*], Karthik Garimella[a], Xin He[b], Christopher Gill[a], Yevgeniy Vorobeychik[a], Xuan Zhang[a]

[a] Washington University in St. Louis United States
[b] University of Michigan, Ann Arbor United States

ARTICLE INFO

Keywords:
Machine learning
Adversarial examples
Autonomous driving
End-to-end learning
Bayesian optimization

ABSTRACT

Recent advances in machine learning, especially techniques such as deep neural networks, are enabling a range of emerging applications. One such example is autonomous driving, which often relies on deep learning for perception. However, deep learning-based perception has been shown to be vulnerable to a host of subtle adversarial manipulations of images. Nevertheless, the vast majority of such demonstrations focus on perception that is disembodied from end-to-end control. We present novel end-to-end attacks on autonomous driving in simulation, using simple physically realizable attacks: the painting of black lines on the road. These attacks target deep neural network models for end-to-end autonomous driving control. A systematic investigation shows that such attacks are easy to engineer, and we describe scenarios (e.g., right turns) in which they are highly effective. We define several objective functions that quantify the success of an attack and develop techniques based on Bayesian Optimization to efficiently traverse the search space of higher dimensional attacks. Additionally, we define a novel class of *hijacking* attacks, where painting lines on the road cause the driverless car to follow a target path. Through the use of network deconvolution, we provide insights into the successful attacks, which appear to work by mimicking activations of entirely different scenarios. Our code is available on https://github.com/xz-group/AdverseDrive

1. Introduction

With billions of dollars being pumped into autonomous vehicle research to reach Level 5 Autonomy, where vehicles will not require human intervention, safety has become a critical issue [3]. Remarkable advances in deep learning, in turn, suggest such approaches as natural candidates for integration into autonomous control. One way to use deep learning in autonomous driving control is in an end-to-end (e2e) fashion, where learned models directly translate perceptual inputs into control decisions, such as the vehicle's steering angle, throttle and brake. Indeed, recent work demonstrated such approaches to be remarkably successful, particularly when learned to imitate human drivers [4].

Despite the success of deep learning in enabling greater autonomy, a number of parallel efforts also have exhibited concerning fragility of deep learning approaches to small adversarial perturbations of inputs such as images [5,6]. Moreover, such perturbations have been shown to effectively translate to physically realizable attacks on deep models, such as placing stickers on stop signs to cause miscategorization of these as speed limit signs [2]. Fig. 1(a) offers several canonical illustrations.

There is, however, a crucial missing aspect of most adversarial attacks to date: manipulations of the physical environment that have a demonstrable *physical* impact (e.g., a crash). For example, typical attacks consider only prediction error as a measure of outcome and focus either on a static image, or a fixed set of views, without consideration of the dynamics of closed-loop autonomous control. To bridge this gap, our aim is to study *end-to-end* adversarial examples. We require such adversarial examples to: 1) modify the physical environment, 2) be simple to implement, 3) appear unsuspicious, and 4) have a physical impact, such as causing an infraction (lane violation or collision). The existing attacks that introduce carefully engineered manipulations fail the simplicity criterion [5,7], whereas the simpler physical attacks, such as stickers on a stop sign, are evaluated solely on prediction accuracy [2].

The particular class of attacks we systematically study is the painting of black lines on the road, as shown in Fig. 1(b). These are unsuspicious since they are semantically inconsequential (few human drivers would be confused) and are similar to common imperfections observed in the

☆ This research was partially supported by NSF awards CNS-1739643, IIS-1905558 and CNS-1640624, ARO grant W911NF1610069 and MURI grant W911NF1810208.

* Corresponding author.
E-mail addresses: adith@wustl.edu (A. Boloor), kvgarimella@wustl.edu (K. Garimella), xinhe@umich.edu (X. He), cdgill@wustl.edu (C. Gill), yvorobeychik@wustl.edu (Y. Vorobeychik), xuan.zhang@wustl.edu (X. Zhang).

https://doi.org/10.1016/j.sysarc.2020.101766
Received 29 August 2019; Received in revised form 19 February 2020; Accepted 16 March 2020
Available online 4 April 2020
1383-7621/

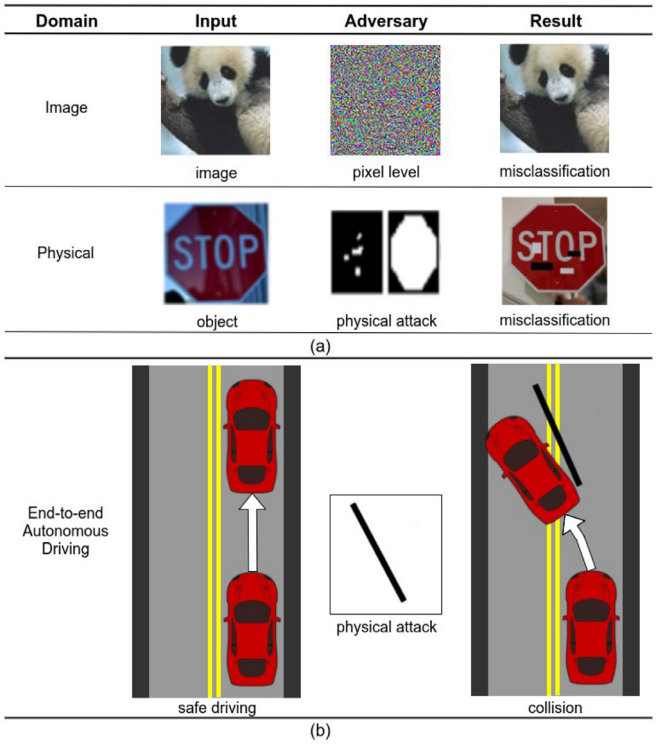

Fig. 1. (a) Existing attacks on machine learning models in the image [1] and the physical domain [2]; (b) conceptual illustration of potential physical attacks in the end-to-end driving domain studied in our work.

real world, such as skid marks or construction markers. Furthermore, we demonstrate a systematic approach for designing such attacks so as to maximize a series of objective functions, and demonstrate actual physical impact (lane violations and crashes) over a variety of scenarios, in the context of end-to-end deep learning-based controllers in the CARLA autonomous driving simulator [8].

We consider scenarios where correct behavior involves turning right, left, and driving straight. Surprisingly, we find that right turns are by far the riskiest, meaning that the right scenario is the easiest to attack; on the other hand, as expected, going straight is comparatively robust to our class of attacks. We use network deconvolution to explore the reasons behind successful attacks. Here, our findings suggest that one of the causes of controller failure is partially mistaking painted lines on the road for a curb or barrier common during left-turn scenarios, thereby causing the vehicle to steer sharply left when it would otherwise turn right. By increasing the dimensionality of our attack space and using a more efficient Bayesian optimization strategy, we are able to find successful attacks even for cases where the driving agent needs to go straight. Our final contribution is a demonstration of novel *hijacking* attacks, where painting black lines on the road causes the car to follow a target path, even when it is quite different from the correct route (e.g., causing the car to turn left instead of right).

This paper is an extension our previous work [9], with the key additions of new objective functions, a new optimization strategy, Bayesian Optimization, and a new type of adversary in the form of hijacking self-driving models. In this paper, we first talk about relevant prior work on deep neural networks, adversarial machine learning in the context of autonomous vehicles, in Section 2. Then in Section 3 we define the problem statement and present several objective functions that mathematically represent the problem statement. In Section 4, we introduce some optimization strategies. In Section 5, we discuss our experimental setup including our adversary generation library and simulation pipeline. Section 6 shows how we were able to successfully generate adversaries against e2e models, and presents a new form of attack, dubbed the hijacking attack where we control the route of the e2e model.

2. Related work

2.1. Deep neural networks for perception and control

Neural Networks (NN) are machine learning models that consist of multiple layers of neurons, where each neuron implements a simple non-linear function (such as a sigmoid function), and where the output is some prediction. Deep Neural Networks (DNNs) are neural networks with more than two layers of neurons, and have increasingly become the state-of-the-art approach for a host of vision based perception problems in the context of autonomous vehicles. Deep convolutional neural networks have been used to detect pedestrians, vehicles and other objects that could serve as obstacles on an autonomous vehicle's path [10–15]. These networks have been trained on large image datasets such as Imagenet [16] and KITTI [17] for detection with nearly human level accuracy. DNNs, along with traditional computer vision practices have been used extensively for lane detection, which is a key part of the self-driving pipeline [18–21]. Furthermore, DNN models have been created for the image segmentation task where the camera images are segmented into different classes such as roads, vehicles, pedestrians, traffic lights, and other hazards [22–25]. Rather than traditional depth estimation algorithms which use stereo images or LiDAR point clouds, DNNs have been used to estimate depth using just single images as input [26–29]. This is an important component of perception in self-driving vehicles so that distances to other vehicles and obstacles can be estimated.

2.2. End-to-end self-driving

While these perception modules are used in various stages of self-driving stacks, end-to-end driving models are capable of directly learning driving decisions from camera images. End-to-end (e2e) learning models for self-driving are comprised of a DNN that accept raw input data like camera images and directly calculate the desired output such as steering angle, throttle, and brake. Rather than explicitly decomposing a complex problem into its constituent parts and solving them separately, e2e self-driving models directly generate driving decisions from a set of inputs. This is achieved by applying gradient-based learning methods to the entire e2e neural architecture. End-to-end models have been shown to have good performance when learning lane-following tasks; one such example is the Autonomous Land Vehicle In a Neural Network model (ALVINN), a 3-layer neural network which took as input a camera image and laser range finder value to output a steering direction in order to follow the road [30]. More recently, e2e learning models driven by Convolutional Neural Networks (CNN) which learn using online imitation learning policies have been shown to be successful in learning off-road driving policies [31]. Previous research has also shown that e2e learning models can be extended to not only make driving decisions but also jointly estimate localization for a fixed environment [32]. In addition to CNN-based e2e models, e2e Long Short Term Memory (LSTM) networks, a form of Recurrent Neural Networks (RNN), have been able to train from only a front camera image in order to predict longitudinal control (i.e. the speed of the autonomous vehicle) [33]. More recently, e2e learning has shown promise in multi-modal learning in which both the driving decision and predicted speed of vehicle are learned simultaneously [34]. Self driving simulators such as CARLA [8] have accelerated the development of research in multi-modal e2e models. For example, several types of multi-modal e2e models have been developed within the CARLA simulator which include models trained from RGB images as well as RGB + Depth (RGBD) images [35]. In contrast to e2e models, self-driving stacks such as Apollo [36] and Autoware [37] decompose the autonomous driving problem into several sub-modules and solve each component individually. Despite having complete autonomous driving

stacks which include trained DNN models for perception, a series of real-world crashes involving autonomous vehicles demonstrate the stakes, and some of the existing limitations of the technology [38–41].

2.3. Attacks on autonomous vehicles

Adversarial examples (also called attacks and adversaries) [5,42–44] are deliberately calculated perturbations to the input which result in an error in the output from a trained DNN model. The idea of using adversarial examples against static image classification models demonstrated that DNNs are highly susceptible to carefully designed pixel-level adversarial perturbations [5,7,45]. More recently, adversarial attacks have been implemented in the physical domain [2,6,46], such as adding stickers to a stop sign that result in misclassification [2]. Additionally, it has been shown that state-of-the-art autonomous driving stacks such as Apollo [36] which rely upon LiDARs are susceptible to physically realizable attacks. In particular, carefully engineered 3D physical objects have been constructed and tested both in simulation and the real-word that remain undetected by Apollo's perception module [47]. Moreover, LiDAR spoofing attacks have been shown to fool the Apollo perception stack to detect a fake object in front of the vehicle thus affecting the planning component. [48,49]. The camera-based object detection components of these driving stacks have also been shown to be susceptible to physical adversaries [2,50]. Recently, researchers have briefly demonstrated that placing stickers on the road can make the Tesla autopilot perceive a lane marker when it does not exist [51].

In this work, we focus on attacking vision based end-to-end self-driving models such as the Imitation Learning and Reinforcement Learning models [8] using physical adversaries.

3. Modeling framework

In this paper, we focus on exploring the influence of a physical adversary that successfully subverts RGB camera-based e2e driving models. We define physical adversarial examples as attacks that are physically realizable in the real world. For example, deliberately painted shapes on the road or on stop signs would be classified as physically realizable. Fig. 1(b) displays the conceptual view of such an attack involving painting black lines. We define our adversarial examples as *patterns*. To create an adversarial example that forces the e2e model to crash the vehicle, we need to choose the parameters of the *pattern*'s shape that maximize the objective functions that we present. This may cause the vehicle to veer into the wrong lane or go offroad, which we characterize as a successful attack. Conventional gradient-based attack techniques are not directly applicable, since we need to run simulations (using the CARLA autonomous driving simulator) both to implement an attack pattern, and to evaluate the end-to-end autonomous driving agent's performance.

At the high level, our goal is to paint a pattern (such as a black line) somewhere on the road to cause a crash. We formalize such attacks in terms of optimizing an objective function that measures the success of the attack pattern at causing driving infractions. Since driving infractions themselves are difficult to optimize because of discontinuity in the objective (infraction either occurs, or not), one of our goals it to identify a high-quality proxy objective. Moreover, since the problem is dynamic, we must consider the impact of the object we paint on the road over a sequence of frames that capture the road, along with this pattern, as the vehicle moves towards and, eventually, over the modified road segment. Crucially, we modify the road itself, which is subsequently captured in vision, digitized, and used as input into the e2e model's controller.

To formalize, we now introduce some notation. Let δ refer to the pattern painted on the road, and let l denote the position on the road where we place the pattern. We use L to denote the set of feasible locations at which we can position the adversarial pattern δ, and S the set of possible patterns (along with associated modifications; in our case, we consider either a single black line, or a pair of black lines, with modifications involving, for example, the distance between the lines, and their rotation angles). Let a_l be the state of the road at position l, and $a_l + \delta$ then becomes the state of the road at this same position when the pattern δ is added to it. The state of the road at position l is captured by the vehicle's vision system when it comes into view; we denote the frame at which this location initially comes into view by F_l, and let Δ be the number of frames over which the road in position l is visible to the vehicle's vision system. Given the road state a_l at position l, the digital view of it in frame F is denoted by $I_F(a_l)$ or simply I_F. Finally, we let $\theta_F = g_{sa}(I_F)$ denote the predicted steering angle given observed digital image corresponding to frame F. With this formalism established, we introduce several candidates for a proxy objective function that would quantify the success of an attack.

3.1. Candidate objective functions

3.1.1. Steering angle summations

First, we denote the vector of predicted steering angles during an episode *with an attack* δ starting from frame F_l to frame $F_{l+\Delta}$ as:

$$\vec{\Theta}_\delta = [\theta_{F_l}, \theta_{F_{l+1}}, \cdots, \theta_{F_{l+\Delta}}] \tag{1}$$

We define two objective functions as:

$$\text{Collide Right} : \max_{l,\delta} \sum_{i=0}^{\Delta} \vec{\Theta}_{\delta_i} \tag{2a}$$

$$\text{Collide Left} : \min_{l,\delta} \sum_{i=0}^{\Delta} \vec{\Theta}_{\delta_i} \tag{2b}$$

$$\text{subject to} : \quad l \in L, \quad \delta \in S. \tag{2c}$$

Eq. (2a) says that to optimize an attack that causes the vehicle to veer off towards the right and collide, we need to maximize the sum of steering angles for that particular experiment for the frames in which the pattern is in view. And similarly in Eq. (2b), we need to minimize the steering sum, to make the vehicle veer left. We convert Eq. (2b) to a maximization problem for consistency in our search procedures that we will describe. Using Eq. 2 as the objective function allows us to have control over which direction we would like the car to crash. The following two metrics, the absolute steering angle difference and path deviation, lose this ability to distinguish direction-based attacks, since they are essentially L-1 and L-2 norms.

3.1.2. Absolute steering angle differences

Again, let's denote the predicted steering angles *with an attack* δ over the frames F_l to $F_{l+\Delta}$ as $\vec{\Theta}_\delta$ as shown in Eq. (1). Now, let's denote the predicted steering angles *without any attack* over the same frames as $\vec{\Theta}_{\text{baseline}}$. This represents an episode where no attack is added to the road (we refer to this as the baseline run) and the car travels the intended path with minimal infractions. We can now define our second candidate metric as:

$$\max_{l,\delta} ||\vec{\Theta}_\delta - \vec{\Theta}_{\text{baseline}}||_1 \tag{3a}$$

$$\text{subject to} : \quad l \in L, \quad \delta \in S. \tag{3b}$$

Eq. (3a) optimizes an attack over the frames Δ that cause the largest absolute deviation in predicted steering angles with respect to the predicted steering angles when no pattern has been added to the road.

3.1.3. Path deviation

First denote the (x, y) position of the agent from frames F_l to $F_{l+\Delta}$ *with an attack* δ as:

$$\vec{p}_\delta = [(x_l, y_l), (x_{l+1}, y_{l+1}), \cdots, (x_{l+\Delta}, y_{l+\Delta})] \tag{4}$$

Define $\vec{p}_{\text{baseline}}$ as the position of the agent *with no attack* added to the road over the same frames (the baseline run). We can optimize the path

deviation from the baseline path:

$$\max_{l,\delta} ||\vec{p}_{\delta} - \vec{p}_{\text{baseline}}||_2 \tag{5a}$$

$$\text{subject to}: \quad l \in L, \quad \delta \in S. \tag{5b}$$

Similar to Eq. (3a), we can use this metric to optimize deviation from the baseline route, except we are now attacking the position of the vehicle which is directly influenced by the outputs of the e2e models.

4. Approaches for generating adversaries

We now describe our approaches for computing adversarial patterns or, equivalently, optimizing the objective functions defined above.

4.1. Random and grid search

Each *pattern* we generate (labeled earlier as δ) can be described by a set of parameters such as length, width, and rotation angle with respect to the road. Two naive methods of finding successful attacks would be to generate a pattern through either a random or grid search (using a coarse grid) and evaluate this pattern using one of the above objective functions. Algorithm 1 shows this setup. The function *RunScenario()*

Algorithm 1 Adversary search algorithm.

Require: Strategy ∈ Random, Grid
  $i \leftarrow 0$
  MetricsList ← [ ]
  loop
    $\delta_i \leftarrow$ GenerateAttack(Strategy)
    results ← RunScenario(δ_i)
    $y_i \leftarrow$ CalculateObjectiveFunction(results)
    MetricsList.append(y_i)
    $i \leftarrow i + 1$
  end loop
  return arg max MetricsList

runs the simulation and returns data such as vehicle speed, predicted acceleration, GPS position, and steering angle. We use these results to calculate one of the objective functions (*CalculateObjectiveFunction()*). As our goal is to maximize this metric, we use *MetricsList* to store the results of the objective function at each iteration. Finally, we return the parameters that maximized our objective function.

4.2. Bayesian optimization search policy

Algorithm 1 works well when the number of parameters for δ are relatively small. For a larger pattern space, and to enable us to explore the space more finely, we turn to Bayesian Optimization, which is designed for optimizing an objective function that is expensive to query without requiring gradient information [52]. It has been shown that Bayesian Optimization (BayesOpt) can be useful for optimizing expensive functions in various domains such as hyper-parameter tuning, reinforcement learning, and sensor calibration [53–56]. In our case, since we use an autonomous driving simulator, it is *expensive* to run a simulation with a generated attack in order to find, for example, the sum of steering angles as shown in Eq. (2). On average, one episode takes between 20 to 40 seconds depending upon the scenario; consequently, it is important for the optimization to sample efficiently.

At the high level, our goal is to generate physical adversaries that successfully attack e2e autonomous driving models, where a successful attack can be quantified as trying to maximize some objective function $f(\delta)$. Our goal, therefore, is to find a physical attack, δ^*, such that:

$$\delta^* = \arg\max_{\delta} f(\delta), \tag{6}$$

where $\delta^* \in \mathcal{R}^d$ and d is the number of parameters of the physical attack. We first assume that the objective f can be represented by a Gaussian Process, which we denote by $GP(f, \mu(\delta), k(\delta, \delta'))$ with a mean function of $\mu(\delta)$ and a covariance function $k(\delta, \delta')$ [57]. We assume the prior mean function to be $\mu(\delta) = \mathbf{0}$ and the covariance function to be the Matérn 5/2 kernel:

$$k(\delta, \delta') = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right), \tag{7}$$

where r is the Euclidean distance between the two input points, $||\delta - \delta'||_2$, and ℓ is a scaling factor optimized during simulation run-time. Let's suppose that we have already generated several adversaries and evaluated our objective function f for each of these adversaries. We can denote this dataset as $D = \{(\delta_1, y_1), \cdots, (\delta_{n-1}, y_{n-1})\}$. Therefore, if we would like to sample our function f at some point along the input space δ, we would obtain some posterior mean value $\mu_{f|D}(\delta)$ along with a posterior confidence or standard deviation value of $\sigma_{f|D}(\delta)$. As noted earlier, our objective function f is expensive to query. When we use Bayesian optimization to find the parameters that define our next adversary δ_n, we instead maximize a proxy function known as the acquisition function, $u(\delta)$. Compared to the objective function, it is trivial to maximize the acquisition function using an optimizer such as the L-BFGS-B algorithm with a number of restarts to avoid local minima. In our case, we utilize the Expected Improvement (EI) acquisition function. Given our dataset, D, we first let $y_{\max}$ be the highest objective function value we have seen so far. The EI can be evaluated at some point δ as:

$$u(\delta) = \mathrm{E}[\max(0, f(\delta) - y_{\max})]. \tag{8}$$

Given the properties of a Gaussian Process, this can be written in closed form as follows:

$$z = \frac{\mu_{f|D}(\delta) - y_{\max}}{\sigma_{f|D}(\delta)}; \tag{9}$$

$$u(\delta) = (\mu_{f|D}(\delta) - y_{\max})\Phi(z) + \sigma_{f|D}(\delta)\phi(z), \tag{10}$$

where Φ and ϕ are the cumulative and probability distribution functions of the Gaussian distribution, respectively. Effectively, the first term in the above acquisition function leads to exploiting information from previously generated adversaries to generate parameters for δ_n while the second term prefers exploring the input space of the adversary parameters. Given this setup, Algorithm 2 presents a Bayesian Optimization approach for generating and searching for adversarial patterns.

Algorithm 2 Bayesian adversary search algorithm.

  $i \leftarrow 0$
  MetricsList ← [ ]
  loop
    $\delta_i \leftarrow \arg\max u(\delta)$
    results ← RunScenario(δ_i)
    $y_i \leftarrow$ CalculateObjectiveFunction(results)
    MetricsList.append(y_i)
    Update Gaussian Process and D with (δ_i, y_i)
    $i \leftarrow i + 1$
  end loop
  return arg max MetricsList

In this algorithm, the Gaussian process is updated in each iteration, and the acquisition function reflects those changes. An initial warm-up phase where the adversary parameters are chosen at random and the simulation is queried for the objective function is used for hyper-parameter tuning.

While Bayesian Optimization has been shown to be an efficient search policy, it is best suited to search spaces with limited dimensionality, typically less than 20 bounded parameters [58]. Our experiments described in Section 5 contains a search space of 4 bounded parameters, a dimensionality sufficiently small for Bayesian Optimization to be effective. In general, our methodology can be applied to vision based e2e

models since the camera input has a direct effect on the objective functions described in Section 3.1. However, in the context of autonomous driving stacks such as Apollo [36] and Autoware [37] in which the objective functions are influenced by several perception modules (e.g. camera, LiDAR, Radar), our adversary generation method would need to be modified to directly influence all modules.

5. Experimental methodology

This section introduces the various building blocks that we use to perform our experiments. Fig. 2 shows the overall architecture of our experimentation method, including the CARLA simulator block, the python client block, and how they communicate with each other to generate and test the attack patterns on the simulator.

5.1. Autonomous vehicle simulator

Autonomous driving simulators are often used to test autonomous vehicles for the sake of efficiency and safety [59–62]. After testing popular autonomous simulators [36,63–65], we choose to run our experiments on the CARLA [8] (CAR Learning to Act) autonomous vehicle simulator, due to its feature-set and ease of source code modification. With Unreal Engine 4 [66] as its backend, CARLA has sufficient flexibility to create realistic simulated environments, with a robust physics engine, lifelike lighting, 3D objects including roads, buildings, traffic signs, vehicles and pedestrians. Fig. 2 shows how the simulator looks in the third person view. It allows us to acquire sensor data like the camera image for each frame (camera view), vehicle measurements (speed, throttle, steering angle and brake) and other environmental metrics like how the vehicle interacts with the environment in the form of infractions and collisions. Since we use e2e models that use only the RGB camera, we disable the LiDAR (Light Detection And Ranging), semantic segmentation, and depth cameras. Steering angle, throttle and brake parameters are the primary control parameters to drive the vehicle in the simulation. CARLA (v0.8.2) comes with two maps: a large training map and a smaller testing map which are used for training and testing the e2e models respectively. CARLA also allows the user to run experiments under various weather conditions like sunset, overcast and rain, which are determined by the client input. To keep consistent frame rate and execution time, we run CARLA using a fixed time-step.

5.2. End-to-end driving models

The CARLA simulator comes with two trained end-to-end models: Conditional Imitation Learning (IL) [67] and Reinforcement Learning (RL) [8]. Their commonality ends at using the camera image as the input to produce output controls that include steering angle, acceleration, and brake. The IL model uses a trained network which consists of demonstrations of human driving on the simulator. In other words, the IL model tries to mimic the actions of the expert from whom it was trained [68]. The IL model's structure comprises of a conditional, branched neural architecture model where the *conditional* part is a high-level command given by the CARLA simulator at each frame. This high-level command can be *left, right or straight at an intersection, and lane follow when not at an intersection*. At each frame, the image, current speed, and high-level command are used as inputs to the branched IL network to directly output the controls of the vehicle. Therefore, each branch is allocated a sub-task within the driving problem (making a decision to cross an intersection or following the current lane). The RL model uses a trained deep network based on a rewards system, provided by the environment based on the corresponding actions, without the aid of human drivers. More specifically, for RL, the asynchronous advantage actor-critic (A3C) algorithm was used. It is worth mentioning that the IL model performed better than the RL model in untrained (test) scenarios [8]. Because of this, we focus our research primarily on attacking the IL model.

5.3. Physical adversary generation

5.3.1. Unreal engine

To generate physically realizable adversaries in a systematic manner, we modify CARLA's source code. The CARLA simulator (v0.8.2) does not allow spawning of objects into the scene which do not already exist in the CARLA blueprint library (which includes models of vehicles, pedestrians, and props). With the Unreal Engine 4 (UE4), we create a new *Adversarial Plane Blueprint*, which is a 200 × 200 pixel plane or canvas with a dynamic UE4 material, which we can overlay on desired portions of the road. The key attribute of this blueprint is that it reads a generated attack image (a.png file) and places it within CARLA in real time. Hence this blueprint has the ability to continuously read an image via an HTTP server. The canvas allows the use of images with an alpha channel which allows attacks which are partly transparent, like the one shown in Fig. 3. Then, we clone the two maps that are provided by CARLA and choose regions of interest within each of them where attacks

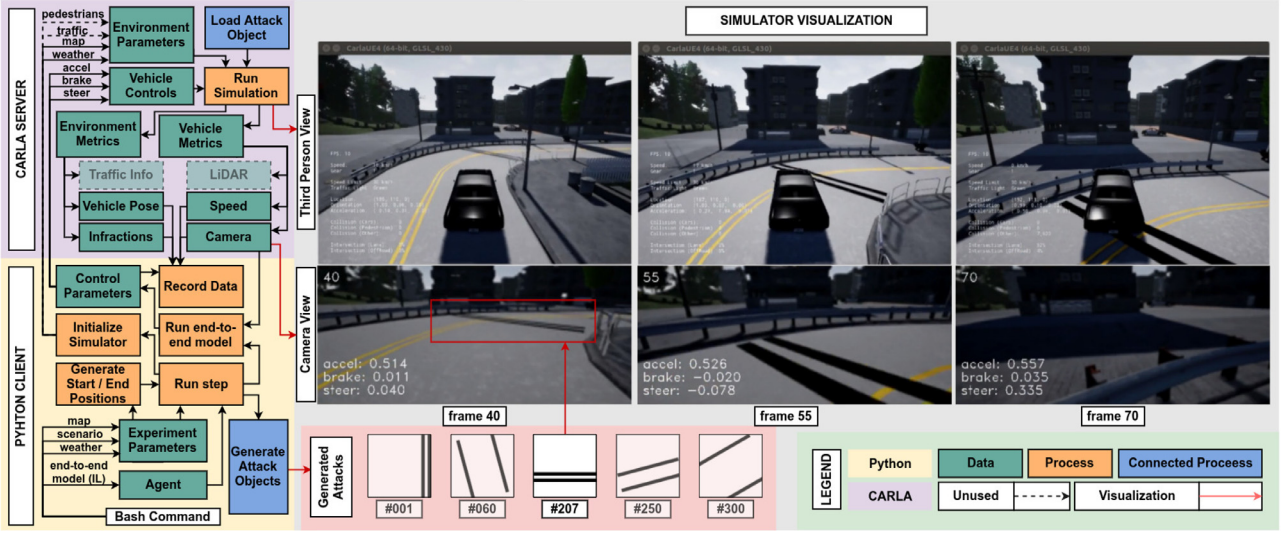

Fig. 2. Architecture overview of our simulation infrastructure including the interfaces between the CARLA simulator and the pattern generator scripts. Visualization of the camera and the third person views from one attack episode are also shown.

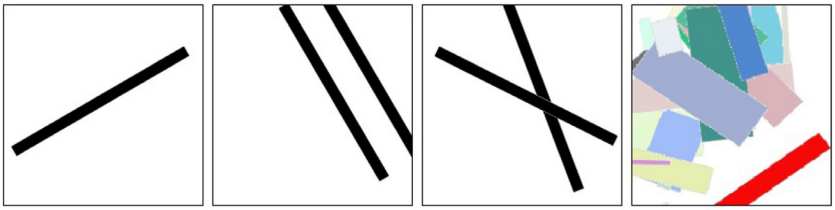

Fig. 3. Attack Generator Capabilities. (a) shows the most basic attack which is a single line. (b) and (c) show attacks using two lines, but (b) has a constraint that the lines always need to be parallel, (d) shows the ability of the generator to generate N number of lines with various shapes and color.

Table 1
Different types of attacks and their respective parameters and constraints. var - variable, const - constant, NA - Not Applicable

	Attack Type			
params	Single Line	Double Line	Two Lines	N-Lines
# lines	1	2	2	N
position	var	var	var	var
rotation	var	var	var	var
length	const	const	const	var
width	const	const	const	var
gap	NA	var	NA	NA
color	const	const	const	var
opacity	const	const	const	var
dimensions	2	3	4	N × 6

spawn. Some interesting regions are at the turns and intersections. We place the *Adversarial Plane Blueprint* canvas in each of these locations. When CARLA runs, an image found on the HTTP server gets overlaid on each canvas. Finally, we compile and package this modified version of CARLA. Hence we are able to place physical attacks within the CARLA simulator.

5.3.2. Pattern generator library

We built a pattern generator that creates different kinds of shapes as shown in Fig. 3 using the pattern parameters (Table 1). For the pattern generator, we explore parameters like the position, width, and rotation of the line(s). We sweep the position and rotation from 0 to 200 pixels and 0–180 degrees respectively to generate variations of attacks. Similarly, we create a more advanced pattern which involves two parallel black lines called the *double-line* pattern as described in Table 1. It comprises of the previous parameters, namely, position, rotation, and width, with the addition of a new gap parameter which is the distance between the two parallel lines. Lastly, we remove the parallel constraint on double lines to increase the search space of the attacks while preserving simplicity. Fig. 2 shows some examples of the generated double line patterns which can be seen overlaid on the road in frames 55 and 70.

Additionally, our library has the ability to read a dictionary object containing the number of lines, the parameters (position, rotation, width, length and color) for each line, and produce a corresponding attack pattern as shown in Fig. 3 (d). Once the pattern is generated, it is read via the HTTP server and is placed within the Carla simulator.

5.3.3. OpenAI-gym environment for carla

Since CARLA runs nearly in real-time, experiments take a long time to run. In order to efficiently run simulations with our desired parameters, we convert the CARLA setup to an OpenAI-Gym environment [69]. While the OpenAI-Gym framework is primarily used for training reinforcement learning models, we find the format helpful as we are able to easily run the simulator with a set of initial parameters like the task (straight, right, left), the map, the scene, the end-to-end model and the desired output metric (eg. average infraction percent for that episode). With this set up, we are able to use an optimizer to generate an attack with a set of defined constraints, run an episode and get the resulting output metric.

5.4. Experiment setup and parallelism

To ensure a broad scope to test the effectiveness of the different attacks in various settings, we conduct experiments by changing various environment parameters like the maps (training map and testing map), scenes, weather (clear sky, rain, and sunset), driving scenarios (straight road, right turn, and left turn), e2e models (IL and RL) and the entire search space for the patterns. Here, we describe the six available driving scenarios for CARLA:

1. *Right Turn*: the agent follows a lane that smoothly turns 90 degrees towards the right.
2. *Left Turn*: the agent follows a lane that smoothly turns 90 degrees towards the left.
3. *Straight Road*: the agent follows a straight path.
4. *Right Intersection*: the agent takes a right turn at an intersection.
5. *Left Intersection*: the agent takes a left turn at an intersection.
6. *Straight Intersection*: the agent navigates straight through intersection.

We choose the baseline scenarios (no attack) where the e2e models drive the vehicle with minimal infractions. We run the experiments at 10 frames per second (fps) and collect the following data for each camera frame (a typical experiment takes between 60 to 100 frames to run): camera image from the mounted RGB camera, vehicle speed, predicted acceleration, steering and brake, percentage of vehicle in the wrong lane, percentage of vehicle on the sidewalk (offroad), GPS position of the vehicle, and collision intensity. Fig. 2 shows this dataflow which is sufficient to assess the ramifications of a particular attack in an experiment.

To search the design space thoroughly, we build a CARLA docker which allows us to run as many as 16 CARLA instances simultaneously, spread over 8 RTX GPUs [70].

6. Experimental results

Through experimentation, we demonstrate the existence of conspicuous physical adversaries that successfully break the e2e driving models. These adversaries do not need to be subtle or sophisticated modifications to the scene. Although they can be distinguished and ignored by humans drivers with ease, they effectively cause serious traffic infractions against the e2e driving models we evaluate.

6.1. Simple physical adversarial examples

6.1.1. Effectiveness of attacks

To begin, we generate two types of adversarial patterns: single line (with varying positions and rotation angles), and double lines (with varying positions, rotation angles, and distance between the lines). In Fig. 4(a), we define different safety regions of the road in ascending order of risk. We start with the vehicle's own lane (safe region), the opposite lane (unsafe), offroad/sidewalk (dangerous) and regions of collisions (very dangerous) past the offroad region. Fig. 4(b)(c)(d)(e) shows that by sweeping through the three scenarios (straight road driving, right turn driving, left turn driving) with the single and double line patterns, for both the training map and testing maps, we see that some pat-

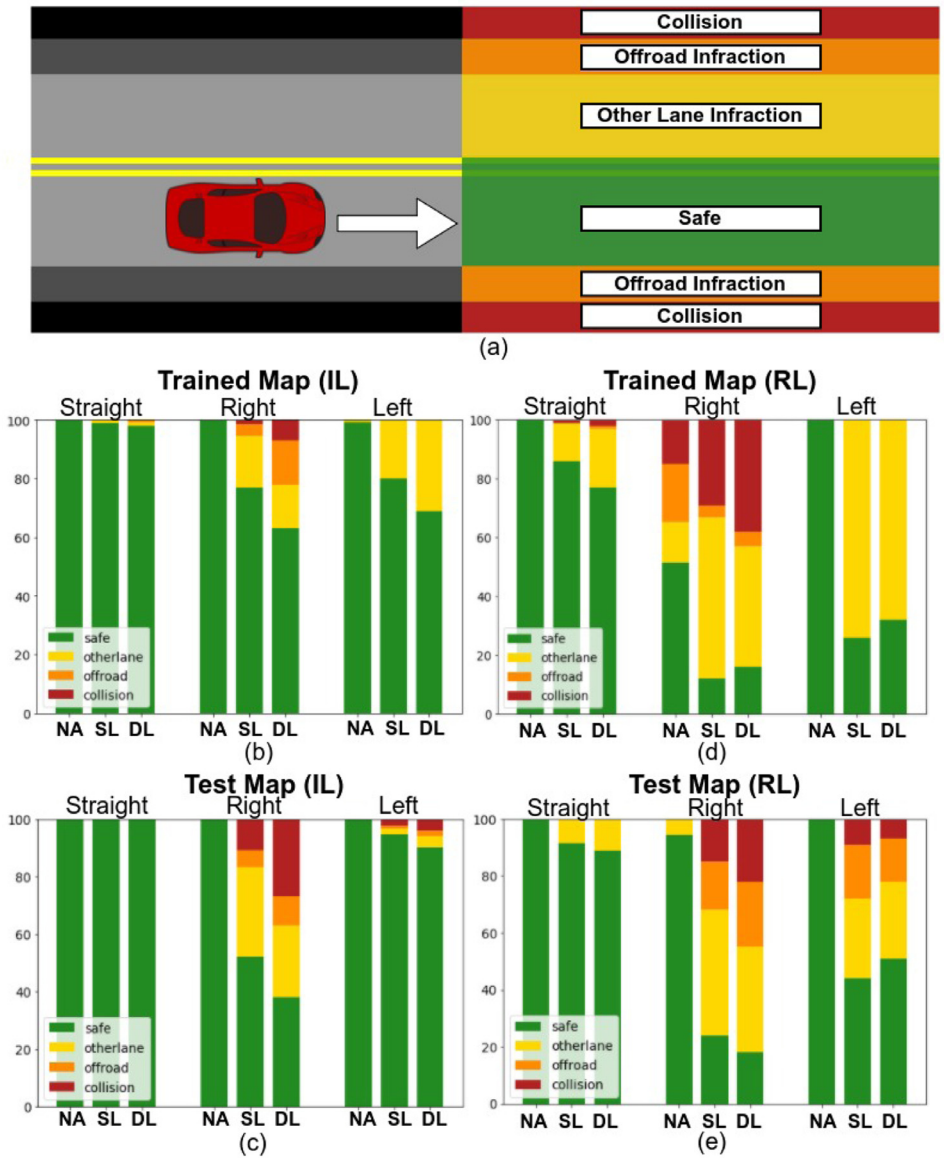

Fig. 4. Comparison of the infractions caused by different patterns. (a) Driving Infraction regions; (b)(c) Infraction percentages for IL; (d)(e) Infraction percentages for RL; NA - No Attack, SL - Single Line pattern, DL - Double Lines pattern; Straight - Straight Road Driving, Right - Right Turn Driving, Left - Left Turn Driving

terns cause infractions. Here we use a naive grid search approach to traverse the search space with the *Steering Sum optimization metric* defined in Eq. (2a). First, we observe the transfer-ability of adversaries since some of our generated adversarial examples cause both IL (Fig. 4(b)) and RL (Fig. 4(d)) models to produce infractions. Second, attacks are more successful against the RL model than the IL model. Additionally, we notice that the double line adversarial examples cause more severe infractions than their single line counterparts. Lastly, we observe that *Straight Road Driving* and *Left Turn Driving* are more resilient to attacks that cause stronger infractions.

6.1.2. Analysis of attack objectives

To find the optimal adversary which produce infractions like collisions for the case of *Right Turn Driving* scenario, the optimizer has to find a pattern that maximizes the first candidate objective function: the sum of steering angles as hypothesized in Eq. (2). A positive steering angle denotes steering towards the right and a negative steering angle implies steering towards the left. Fig. 5(a)(b) show the sum of steering angles and the sum of infractions respectively, for each of the 375 combinations of double line patterns. The infractions are normalized because collision data is recorded in SI units of intensity $[kg \times m/s]$, whereas the lane infractions are in percentages of the vehicle area in the respective regions. Fig. 5 also shows the three lowest points (minima) for the steering sum and the three highest points (maxima) for the collisions plot. In Fig. 5(c), we use the *argmin* and *argmax* on the set of attacks to observe the shapes of the corresponding adversarial examples for both the steering sum and infraction results. We observe that the *patterns that minimize the sum of steering angle and correspondingly maximize the collision intensity are very similar*. Thus, the objective based on maximizing or minimizing steering angles is clearly yielding valuable information for the underlying optimization problem. However, this does not mean that it's the best objective, among the three choices we considered above. We explore this issue in greater depth in the next subsection, as we move towards studying more complex attacks using Bayesian optimization.

6.2. Exploration of large design spaces

In Fig. 4, we observe that when we switch from a *Single Line* attack (with 2 dimensions) to a *Double Line* attack (with 3 dimensions), in most cases, there is a significant increase in the number of successful attacks. It is reasonable to assume that as we increase the number of degrees of

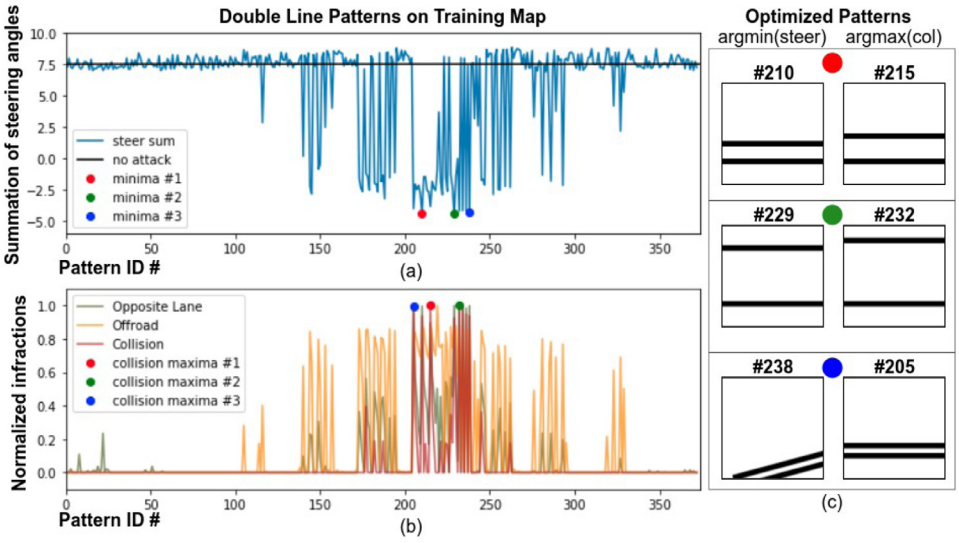

Fig. 5. Adversary against "Right Turn Driving". (a) Adversarial examples significantly changes the steering control. (b) Some patterns cause minor infractions whereas others cause level 3 infractions. (c) The patterns that cause the minimum steering sum and maximum collisions look similar.

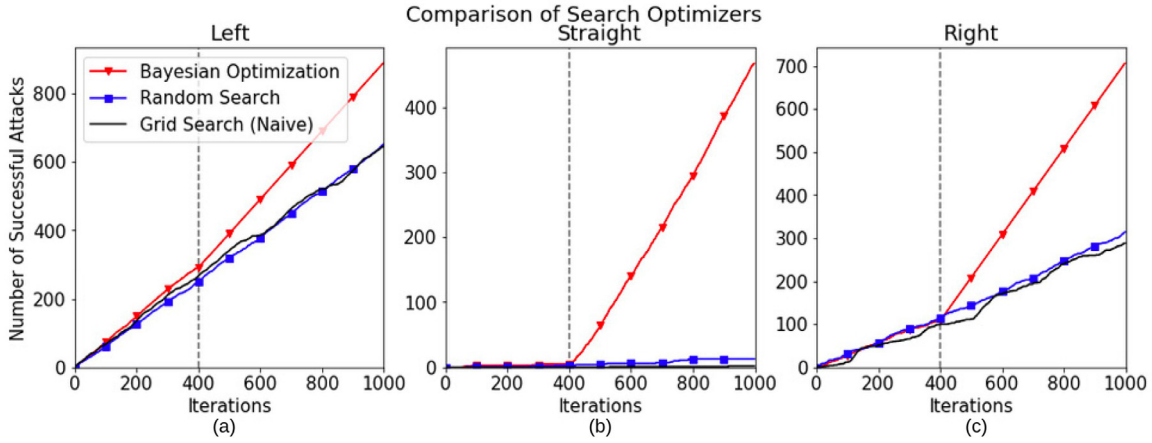

Fig. 6. A comparison of different search algorithms for generating successful attacks. In each driving scenario: Left Turn (a), Straight Road (b), and Right Turn (c) Driving, the Bayesian approach not only finds more unique, successful adversaries in the same number of iterations but also finds these attacks at a faster rate. BayesOpt randomly samples the adversary search space for the first 400 iterations (shown before the dashed line) to tune the hyper-parameters of the kernel function. After these randomly sampled points, BayesOpt utilizes an acquisition function to sample the search space. While a dense grid search would eventually find at the least the same number of attacks as BayesOpt, we constrain our experiments to 1000 iterations given our computational resources.

freedom in the attack, it should be possible to also increase the success rate. We lend further support to this intuition by considering an attack called the *Two Line* attack, shown in Fig. 3(c), with 4 dimensions by removing the constraint that the two lines must be parallel. As shown in Fig. 4, attack success rates increase considerably compared to the more restricted attack.

However, increasing the dimensionality of the attack search space makes grid search impractical. For example, the Single Line attack with grid search requires around 375 iterations to sweep the search space at a 20 pixel resolution. Preserving the same parameter resolution (or precision) would require 1440 iterations for Double Lines, and 12,960 iterations for the Two Line attack. Naive search would require around 45 days to sweep through the search space for a *single scenario* on a modern GPU. Additionally, using a sparser resolution for the attack parameters means that we would not find potential attacks which can only be found at a higher resolution.

We address this issue by adopting the Bayesian Optimization framework (BayesOpt) for identifying attack patterns (introduced in Section 4.2). This requires a change in our search procedure as shown in Algorithm 2. In short, it uses the prior history of the probed search space to suggest the next probing point.

Fig. 6 shows the comparison between the 3 optimization techniques we employ for the straight, left-turn, and right-turn scenarios. We see that for all three cases, BayesOpt outperforms the naive grid search and the random search methods. In Fig. 6, BayesOpt uses 400 initial random points to sample the search space and subsequently samples 600 optimizing points. Hence, we observe that for the first 400 iterations, BayesOpt follows closely with random search, and after probing those initial random points, we observe a significant increase in the number of successful attacks.

Because we observe many more successful attacks against the Left and Right Turn scenarios as compared to the Straight Scenario, Fig. 6 further supports our notion that driving straight is harder to attack as compared to the right and left turn scenarios.

Equipped with BayesOpt, we now systematically evaluate the relative effectiveness of the different objective functions. Table 2 shows the infractions caused by each of the objective functions (path deviation, sum of steering angles, and absolute difference in steering angles with the baseline). For Left Turn, Straight Road, and Right Turn Driving, we list the percentage out of 600 simulation runs using BayesOpt that were safe, incurred collisions, off road infractions, or opposite lane infractions. We observe that the absolute difference in steering angles with

Table 2
Comparison of Candidate Objective Functions as listed in Section 3 (in %). Σ st. angles - sum of steering angles, abs. st. diff. - absolute steering difference

	Left				Straight				Right			
Metric	safe	collision	offroad	opp. lane	safe	collision	offroad	opp. lane	safe	collision	offroad	opp. lane
Σ st. angles	18.2	0	0	81.8	99	0	0	1	72.2	9.5	13.8	24.5
path deviation	64.6	0	0	35.4	23.8	2.5	2.8	76.2	57.2	24.0	28.3	40.2
abs. st. diff.	**0.2**	0	0	**99.8**	**22.7**	**7.5**	**9.3**	**77.3**	**0**	**95.2**	**99.2**	**100**

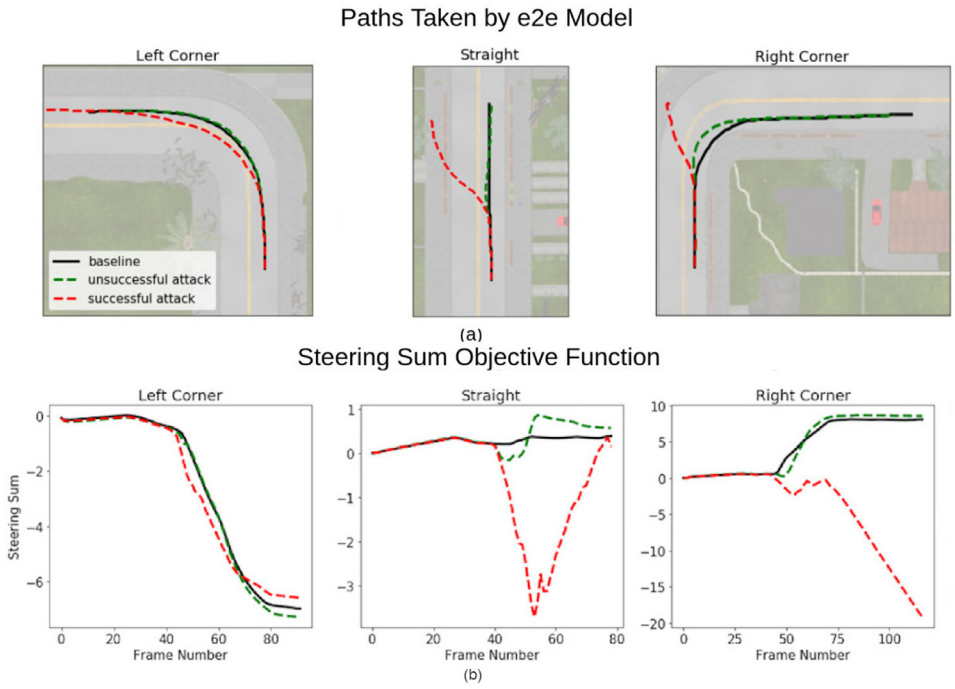

Fig. 7. Paths taken by e2e model in Left Turn, Straight Road, and Right Turn Driving with no attack (baseline), an unsuccessful attack, and a successful attack (a). Cumulative sum of steering angles for each scenario (b). While the successful attack is able to cause the e2e agent to incur an infraction or collision in each scenario, the steering sum metric is unable to capture distinguish between the successful and unsuccessful attack in two of the three scenarios.

respect to the baseline run is the strongest metric when coupled with BayesOpt to discover unique, successful attacks. While the most natural metric would seem to be *steering sum*, it is in practice considerably less effective than maximizing absolute difference in the steering angle. The *path deviation* objective function performs well in *right turn* and *straight scenarios*, but fails to find optimal attacks in the *left turn driving* scenario. Overall it still under-performs when compared to the absolute steering difference objective function.

6.3. Importance of selecting a reliable objective function

In Section 6.2, we evaluate three different objective functions: *path deviation, sum of steering angles* and *absolute steering difference*. We observe that the choice of the right objective function is crucial for success, and this choice is not necessarily obvious.

Most surprisingly, perhaps, we find that the objective that uses the steering angles to guide adversarial example construction is not the best choice, even though it is perhaps the first that comes to mind, and one used in prior work [61]. We now investigate why this choice of the objective can fail.

Fig. 7 shows three driving scenarios (left turn, driving straight, and right turn) respectively. Fig. 7(a) shows the paths taken by the vehicle for 3 cases: a *baseline* case where there is no attack, an *unsuccessful attack* case where an attack pattern does not cause the e2e model to deviate significantly from the *baseline* path, and a *successful attack* case where an attack causes a large deviation resulting in an infraction. Fig. 7(b) shows the sum of steering angles for each of the corresponding cases in Fig. 7(a). Note that for *Left Turn Driving*, we try to maximize Eq. (2a), which is to collide to the right, and for *Straight Driving* and *Right Turn Driving*, we maximize Eq. (2b), which is to collide to the left. For the *right turn driving* scenario, we observe that there is indeed a large difference between the steering sums for a strong attack and a weak attack, but in the other two scenarios, we notice that the baseline, unsuccessful attack and successful attack have very similar steering sums. Hence, the optimizer has a difficult time distinguishing between an unsuccessful and successful attack. In *straight driving scenario*, we see that the steering sum for a successful attack begins increasing and then sharply decreases, even though the vehicle has deviated significantly from the baseline path. This is due to the ability of the IL e2e model to recover in this case, resulting from data augmentation at training time where initial position of the car was randomly perturbed. The *sum of steering angles* objective function is unable to capture this behavior. For the case of *left turn driving*, we discover that the successful attack not only causes a change in steering angle, but also a change in throttle, resulting in the vehicle speeding up and reaching a position further along the baseline path, which opens up new possibilities for generating attacks as well as causing new types of infractions.

The *absolute steering difference* mitigates the above issues by summing up the absolute steering differences between the baseline and attack cases. This allows the objective function to counteract the recovery ability of the e2e models. However, we do lose the ability to directly control the direction towards which we desire the vehicle to crash.

6.4. Vehicle hijacking attacks

Thus far, our exploration of adversarial examples against autonomous driving models focused on causing the car to crash, or cause other infractions. We now explore a different kind of attack: vehicle hijacking. In this attack, the primary purpose is to stealthily lead the car along a target path of the adversary's choice.

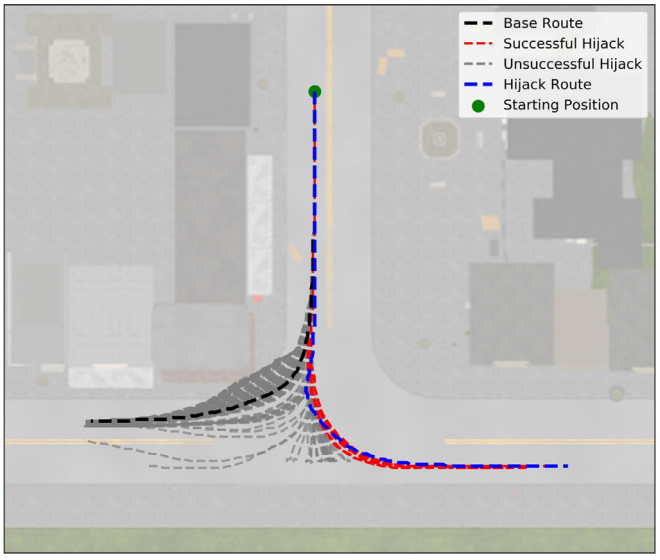

Fig. 8. Illustration of a *hijack attack* where we use an attack to trick the vehicle to deviate from its normal path (base route) to a target hijack route. It demonstrates a successful hijack where we make a vehicle otherwise turning right at an intersection, to turn left.

When attacking the IL model, previous experiments have only targeted the *Lane Follow* branch of this model. Now, we focus our attacks on three different branches of the IL Model: *Right Intersection, Left Intersection,* and *Straight Intersection*. Here, we define a successful attack to be an adversary that 1) causes no infractions or collisions and 2) causes the agent to make a turn chosen by the attacker rather than the ground truth at a particular intersection (e.g. the attacker creates an adversary to make the agent turn left instead of go straight through an intersection). With this definition, an attack that causes the agent to incur an infraction is not considered to be a successful attack. In order to produce such attacks, we modify our experimental setup. After choosing a particular intersection, we run the simulation with no attack to record the baseline steering angles over the course of the episode. The high-level command provided by CARLA directs the agent to take a particular action at that intersection (for example, go straight). We then modify the CARLA high-level command to the direction desired by the attacker (for example, take a right turn). After running the simulation, we store these target steering angles over the entire episode. Finally, we revert the CARLA high-level command to the original command provided to the agent during the baseline simulation run and begin generating attacks at the intersection. We modify our optimization problem to minimize the difference in the steering angles recorded during an episode with an attack ($\vec{\Theta}_\delta$ as defined in 3.1) and the steering angles of the target run (defined as $\vec{\Theta}_{\text{target}}$):

$$\min_{l,\delta} ||\vec{\Theta}_\delta - \vec{\Theta}_{\text{target}}||_1 \tag{11a}$$

$$\text{subject to}: \quad l \in L, \quad \delta \in S. \tag{11b}$$

CARLA (v0.8.2) does not include a four-way intersection in their provided maps, which constrain our experiments to a three-way junction as shown in Fig. 8. Of the six possible hijacking configurations, we are able to generate adversaries that successfully hijacked the car to take a desired route rather than the baseline route for five configurations. For example, Fig. 8 shows the car being hijacked to take a right turn instead of going straight. While we are able to produce attacks that incurred an infraction in each scenario shown in Fig. 8 (the gray paths), these episodes did not count as successful hijacks as the car did not take the target route. Table 3 shows the rate of successful attacks for the six available hijacking scenarios in CARLA v0.8.2. To conclude, we are able to modify our optimization problem and generate adversaries at intersections which caused the agent to take a hijacking route, rather than the intended route.

Table 3
Success rate of Hijacking Attacks for six scenarios.

Hijack Success Rates	% Successful	% Unsuccessful
Straight → Right	14.8	85.2
Straight → Left	0.0	100.0
Left → Straight	23.7	76.3
Left → Right	14.3	85.7
Right → Left	1.4	98.6
Right → Straight	25.9	74.1

6.5. Interpretation of Attacks using DeConvNet

In this section, our goal is to better understand what makes the attacks effective. We begin by quantitatively analyzing the range of parameters of attacks that will generate the most robust attacks in the context of right turns. For simplicity, we analyze the Double Line attack whose parameters include rotation angle, position, and gap size. Fig. 9 shows a histogram of the collision incidence rates versus the pattern IDs, and its corresponding parameters for an experiment with 375 iterations. Fig. 9(b), in particular, shows that some parameters play a stronger role than others in generating a successful attack. For example, in this particular setting Double Line attacks, successful adversaries have a narrow range of rotation angles (90 - 115 degrees). Fig. 9(b) also shows that smaller gap sizes perform slightly better than larger ones.

To better understand the working mechanisms of the successful attack to the underlying imitation learning algorithm, we use network deconvolution, using a state-of-the-art technique, DeConvNet [71]. Specifically, we attach each CONV block (a convolution layer with ReLU and a batch normalizer) to a DeConv counterpart, since the backbone of the imitation learning algorithm is a convolutional neural network which consists of eight CONV blocks for feature extraction and two fully connected (FC) blocks for regression. Each DeConv block uses the same filters, batch norm parameters, and activation functions as the CONV block, except that the operations are reversed. In this paper, DeConvNet is used merely as a probe to the already trained imitation learning network: it provides a continuous path to map high-level feature maps down to the input image. To interpret the network, the imitation learning network first processes the input image and computes the feature maps throughout the network layers. To view selected activations in the feature maps of a layer, other activations are set to zero, and the feature maps backtrack through the rectification, reverse-batch norm, and transpose layers. Then, activations that contribute to the chosen activations in the lower layer are reconstructed. The process is repeated until the input pixel space is reached. Finally, the input pixels which give rise to the activations are visualized. In this experiment, we choose the *top-200* strongest/largest activations in the *fifth* convolution layer and mapped these activations down to the input pixel space for visualization. The reasons behind this choice are twofold: 1) The strongest activations stand out and dominate the decision-making in NNs and the *top-200* activations are sufficient to cover the important activations, and 2) activations of the fifth CONV layer are more representative than other layers, since going deeper would mean that the amount of non-zero activations reduces significantly, which invalidates the deconvolution operations, while shallow layers fail to fully capture the relation between different extracted features.

We conduct a case study to understand why an attack works. Specifically, we take a deeper look inside the imitation network when adversaries are attacking the autonomous driving model for the right turn driving scenario. The baseline case without any attack is depicted in Fig. 10(a) while the one with a successful double-line attack is shown in Fig. 10(b). In the first row of Fig. 10, the input images from the

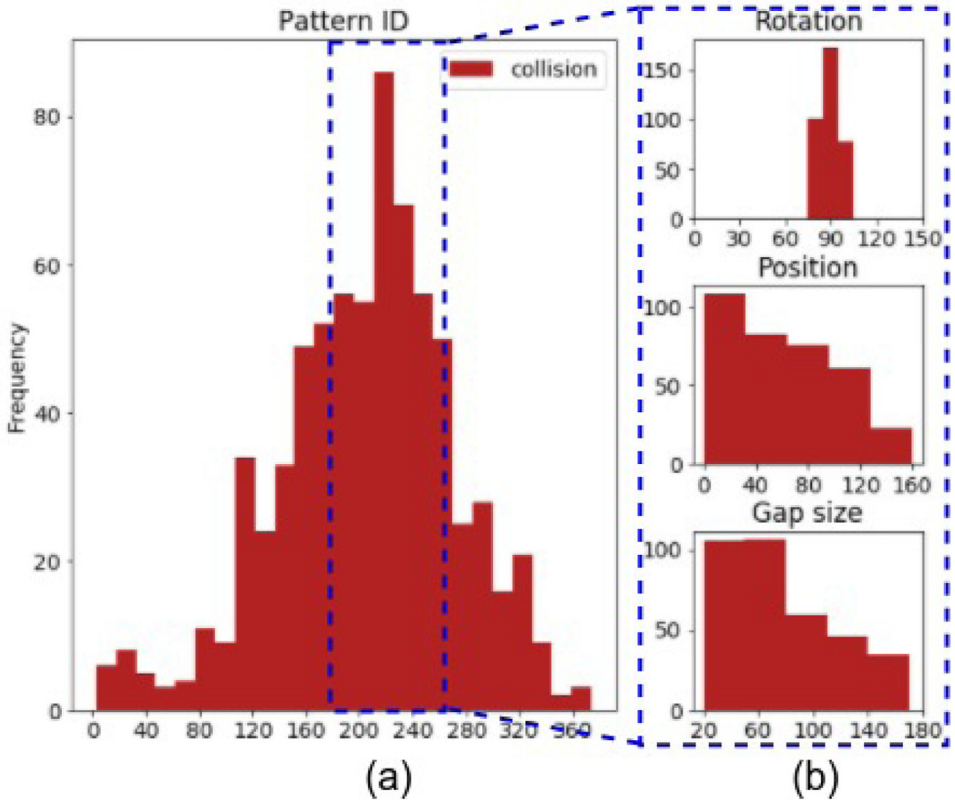

Fig. 9. (a) Histogram showing strong adversaries. (b) Depiction of range of rotation, position and gap parameters for the most robust adversaries.

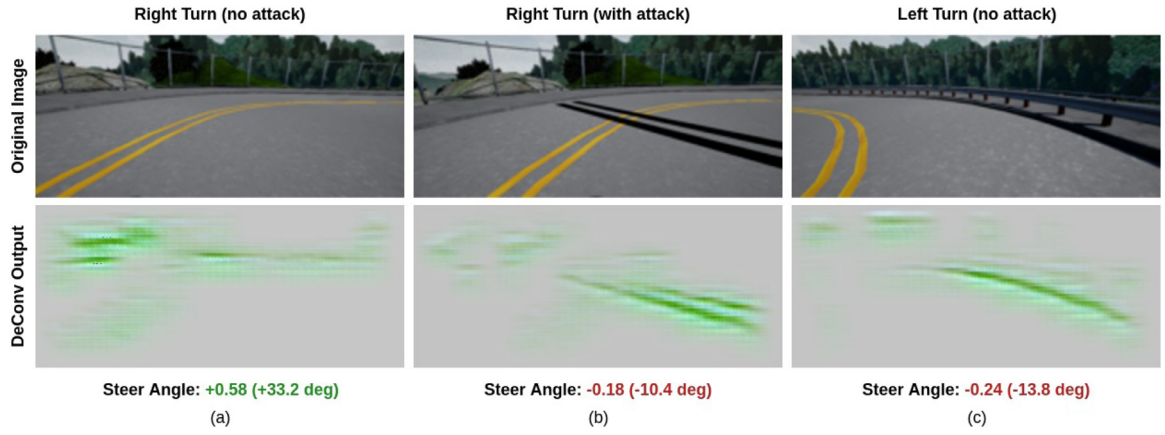

Fig. 10. Attacks against Right Turn Driving: The top row shows the camera input while the bottom deconvolution images show that the reconstructed inputs from the strongest activations determine the steering angle. (a) Right Turn Driving without attack, (b) Right Turn Driving with attack and (c) Left Turn Driving without attack for comparison

front camera mounted on the vehicle are displayed, which are fed to the imitation learning network. In Fig. 10(a), the imitation learning network guides the vehicle to turn right at the corner, as the steering angle output is set to a positive value (steering +0.58). The highlighted green regions in the reconstructed inputs in the corresponding second row show the imitation network makes this steering decision mainly following the curve of the double yellow line. However, when deliberate attack patterns are painted on the road as shown in Fig. 10(b), the imitation network fails to perceive the painted lines which could be easily ignored by a human; instead, the network regards the lines as physical barriers and guides the vehicle to steer left (steering -0.18) to avoid a fictitious collision, leading to an actual collision. The reconstructed image below confirms that the most outstanding features are the painted adversaries instead of the central double yellow lines. We speculate that the vehicle recognizes the adversaries as the road curb. And Fig. 10(c) confirms our speculations. In this case, the vehicle is turning left and the corresponding reconstructed image shows the curb would contribute the strongest activations in the network which will make the steering angle a negative value (steering -0.24) to turn left. The similarity of the reconstructed inputs between cases (b) and (c) suggests that the painted attacks are misrecognized as a curb which leads to an unwise driving decision. To summarize, the deliberate adversaries that mimic important road features are very likely to be able to successfully attack the imitation learning algorithm. This also emphasizes the importance of taking more diverse training samples into consideration when designing autonomous driving techniques. Note that since the imitation learning network makes driving decisions solely based on current camera input, using one frame per case

for visualization is enough to unravel the root causes of an attack's success.

7. Conclusion

In this paper, we develop a versatile modeling framework and simulation infrastructure to study adversarial examples on e2e autonomous driving models. Our model and simulation framework can be applied beyond the scope of this paper, providing useful tools for future research to expose latent flaws in current models with the ultimate goal of improving them. Through comprehensive experiment results, we demonstrate that simple physical adversarial examples that are easily realizable, such as mono-colored single-line and multi-line patterns, not only exist, but can be quite effective under certain driving scenarios, even for models that perform robustly without any attacks. We demonstrate that Bayesian Optimization coupled with a strong objective function is an effective approach to generating devastating adversarial examples. We also show that by modifying the objective function, we are able to hijack a vehicle where we cause the driverless car to deviate from its original route to a route chosen by an attacker. Finally, our analysis using the DeConvNet method offers critical insights to further explore attack generation and defense mechanisms. Our code repository is available at: https://github.com/xz-group/AdverseDrive.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Dr. Ayan Chakrabarti for his advice on matters related to computer vision with this research and Dr. Roman Garnett for his suggestions regarding Bayesian Optimization. We would also like to thank the CARLA team for their technical support regarding the CARLA simulator. This research was partially supported by NSF awards CNS-1739643, IIS-1905558 and CNS-1640624, ARO grant W911NF1610069 and MURI grant W911NF1810208.

References

[1] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2015) arXiv:1412.6572.
[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning models, 2017.
[3] R. Fan, J. Jiao, H. Ye, Y. Yu, I. Pitas, M. Liu, Key ingredients of self-driving cars, 2019.
[4] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L.D. Jackel, U. Muller, Explaining how a deep neural network trained with end-to-end learning steers a car (2017) arXiv:1704.07911.
[5] Y. Vorobeychik, M. Kantarcioglu, Adversarial Machine Learning, Morgan and Claypool, 2018.
[6] T. Dreossi, S. Jha, S.A. Seshia, Semantic adversarial deep learning, CAV, 2018.
[7] N. Papernot, P.D. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, 2016 IEEE European Symposium on Security and Privacy (EuroS&P) (2016) 372–387.
[8] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, V. Koltun, Carla: An open urban driving simulator, CoRL, 2017.
[9] A. Boloor, X. He, C. Gill, Y. Vorobeychik, X. Zhang, Simple physical adversarial examples against end-to-end autonomous driving models, in: 2019 IEEE International Conference on Embedded Software and Systems (ICESS), 2019, pp. 1–7, doi:10.1109/ICESS.2019.8782514.
[10] C. Papageorgiou, T. Poggio, A trainable system for object detection, International journal of computer vision 38 (1) (2000) 15–33.
[11] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[12] G. Prabhakar, B. Kailath, S. Natarajan, R. Kumar, Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving, in: 2017 IEEE Region 10 Symposium (TENSYMP), IEEE, 2017, pp. 1–6.
[13] C. Caraffi, T. Vojíř, J. Trefný, J. Šochman, J. Matas, A system for real-time detection and tracking of vehicles from a single car-mounted camera, in: 2012 15th International IEEE Conference on Intelligent Transportation Systems, 2012, pp. 975–982.
[14] X. Chen, K. Kundu, Y. Zhu, A.G. Berneshawi, H. Ma, S. Fidler, R. Urtasun, 3d object proposals for accurate object class detection, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 424–432.
[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2012) 84–90.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009) 248–255.
[17] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, The International Journal of Robotics Research 32 (11) (2013) 1231–1237.
[18] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, et al., An empirical evaluation of deep learning on highway driving (2015) arXiv preprint arXiv:1504.01716.
[19] J. Li, X. Mei, D. Prokhorov, D. Tao, Deep neural network for structural prediction and lane detection in traffic scene, IEEE transactions on neural networks and learning systems 28 (3) (2016) 690–703.
[20] J. Kim, M. Lee, Robust lane detection based on convolutional neural network and random sample consensus, in: International conference on neural information processing, Springer, 2014, pp. 454–461.
[21] A. Gurghian, T. Koduri, S.V. Bailur, K.J. Carey, V.N. Murali, Deeplanes: End–to-end lane position estimation using deep neural networksa, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 38–45.
[22] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.
[23] D. Levi, N. Garnett, E. Fetaya, I. Herzlyia, Stixelnet: A deep convolutional network for obstacle detection and road segmentation., BMVC, 2015. 109–1
[24] M. Ren, R.S. Zemel, End-to-end instance segmentation with recurrent attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6656–6664.
[25] G.L. Oliveira, W. Burgard, T. Brox, Efficient deep models for monocular road segmentation, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 4885–4891.
[26] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Advances in neural information processing systems, 2014, pp. 2366–2374.
[27] Y. Kuznietsov, J. Stuckler, B. Leibe, Semi-supervised deep learning for monocular depth map prediction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6647–6655.
[28] F. Liu, C. Shen, G. Lin, Deep convolutional neural fields for depth estimation from a single image, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5162–5170.
[29] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, I. Reid, Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 340–349.
[30] D. Pomerleau, ALVINN: an autonomous land vehicle in a neural network, in: D.S. Touretzky (Ed.), Advances in Neural Information Processing Systems 1, [NIPS Conference, Denver, Colorado, USA, 1988], Morgan Kaufmann, 1988, pp. 305–313.
[31] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, B. Boots, Agile autonomous driving using end-to-end deep imitation learning, 2017.
[32] A. Amini, G. Rosman, S. Karaman, D. Rus, Variational end-to-end navigation and localization, 2018.
[33] L. George, T. Buhet, E. Wirbel, G. Le-Gall, X. Perrotton, Imitation learning for end to end vehicle longitudinal control with forward camera, 2018.
[34] Z. Yang, Y. Zhang, J. Yu, J. Cai, J. Luo, End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perception, 2018.
[35] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, A.M. López, Multimodal end-to-end autonomous driving, 2019.
[36] Baidu, Apollo, (http://apollo.auto/).
[37] TierIV, autoware, 2019, (https://www.autoware.ai/).
[38] S. Alvarez, Research group demos why tesla autopilot could crash into a stationary vehicle, 2018, (https://www.teslarati.com/tesla-research-group-autopilot-crash-demo/).
[39] T.S., Why uber's self-driving car killed a pedestrian, 2018, (https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian).
[40] T. Lee, Driverless car from gms cruise and motorcycle collide in san francisco, 2017, (https://arstechnica.com/cars/2017/12/driverless-car-from-gms-cruise-and-motorcycle-collide-in-san-francisco/).
[41] A. Davies, Google's self-driving car caused its first crash, 2016, (https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/).
[42] A. Chernikova, A. Oprea, C. Nita-Rotaru, B. Kim, Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction, 2019.
[43] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, 2018.
[44] D. Lowd, C. Meek, Adversarial learning, KDD, 2005.
[45] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, NIPS, 2014.
[46] J. Lu, H. Sibai, E. Fabry, D.A. Forsyth, No need to worry about adversarial examples in object detection in autonomous vehicles (2017) arXiv:1707.03501.
[47] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q.A. Chen, K. Fu, Z.M. Mao, Adversarial sensor attack on lidar-based perception in autonomous driving (2019) arXiv:1907.06826.

[48] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, B. Li, Adversarial objects against lidar-based autonomous driving systems (2019) arXiv:1907.05418.
[49] H. Shin, D. Kim, Y. Kwon, Y. Kim, Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications, 2017. https://eprint.iacr.org/2017/613
[50] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A.L. Yuille, Adversarial examples for semantic segmentation and object detection (2017) arXiv:1703.08603.
[51] T.K.S. Lab, Tencent keen security lab: Experimental security research of tesla autopilot, 2019, (https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/).
[52] E. Brochu, V.M. Cora, N. de Freitas, A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, 2010.
[53] P.I. Frazier, A tutorial on bayesian optimization, 2018.
[54] M.O. R. Garnett, S. Roberts., Bayesian optimization for sensor set selection, 2010.
[55] J.C. Barsce, J.A. Palombarini, E.C. Martínez, Towards autonomous reinforcement learning: Automatic setting of hyper-parameters using bayesian optimization, 2018.
[56] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M.M.A. Patwary, Prabhat, R.P. Adams, Scalable bayesian optimization using deep neural networks, 2015.
[57] C.E. Rasmussen, Gaussian processes for machine learning, MIT Press, 2006.
[58] R. Moriconi, M.P. Deisenroth, K.S.S. Kumar, High-dimensional bayesian optimization using low-dimensional feature spaces, 2019.
[59] S. Shah, D. Dey, C. Lovett, A. Kapoor, Airsim: High-fidelity visual and physical simulation for autonomous vehicles, FSR, 2017.
[60] H. Fan, F. Zhu, C. Liu, L. Zhang, L. Zhuang, D. Li, W. Zhu, J. Hu, H. Li, Q. Kong, Baidu apollo em motion planner (2018) arXiv:1807.08048.
[61] Y. Tian, K. Pei, S. Jana, B. Ray, Deeptest: Automated testing of deep-neural-network-driven autonomous cars, 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE) (2018) 303–314.
[62] C.E. Tuncali, G. Fainekos, D. Prokhorov, H. Ito, J. Kapinski, Requirements-driven test generation for autonomous vehicles with machine learning components, 2019.
[63] C. Quiter, M. Ernst, deepdrive/deepdrive: 2.0, 2018, (https://doi.org/10.5281/zenodo.1248998). 10.5281/zenodo.1248998
[64] NVIDIA, Driveworks, (https://developer.nvidia.com/drive/drive-software).
[65] Microsoft, Microsoft airsim, 2018, (https://github.com/microsoft/AirSim).
[66] Epic Games Inc., What is unreal engine?, 2019, (https://www.unrealengine.com).
[67] F. Codevilla, M. Miiller, A. López, V. Koltun, A. Dosovitskiy, End-to-end driving via conditional imitation learning, 2018 IEEE International Conference on Robotics and Automation (ICRA) (2018) 1–9.
[68] A. Attia, S. Dayan, Global overview of imitation learning, 2018.
[69] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, Openai gym, 2016.
[70] NVIDIA Corporation, What is geforce rtx?, 2019, (https://www.nvidia.com/en-us/geforce/20-series/rtx/).
[71] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, ECCV, 2014.

Adith Boloor is a PhD student in Computer Science at Washington University in St. Louis. He has a Master's degree in Robotics from Washington University, and a Bachelor's degree in Mechanical Engineering from Purdue University. He has worked on mutli-agent systems, additive manufacturing, humanoid robots, autonomous vehicles and deep learning. His research interests include adversarial machine learning in the context of self-driving vehicles. He was invited to give a talk at CVPR 2019 for his work on creating end-to-end self driving agents.

Karthik Garimella is a MSc student in Computer Engineering at Washington University in St. Louis. Before starting there, he completed his undergraduate degree in Physics from Hendrix College. He has worked as a scientific software developer for several NASA sites, including Oak Ridge National Lab, Goddard Space Flight Center, and the Jet Propulsion Laboratory. His research interests include machine learning and artificial intelligence for autonomous systems.

Xin He (M'17) is a postdoctoral research fellow at the University of Michigan, Ann Arbor. He received the PhD degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2017. His research interests include computer architecture, especially on application specific acceleration, deep learning, neural network accelerator, and approximate computing. He is a member of the IEEE.

Professor Gill's research focuses on assuring properties of real-time and embedded systems in which software complexity, interactions with unpredictable environments, and heterogeneous platforms demand novel solutions that are grounded in sound theory. A major goal of his work is to assure that constraints on timing, memory footprint, fault-tolerance, and other system properties can be met across heterogeneous applications, operating environments and deployment platforms. He has led or contributed to the development, evaluation, and open-source release of numerous real-time systems research platforms and artifacts, including the Kokyu real-time scheduling and dispatching framework that was used in several AFRL and DARPA projects and flight demonstrations; the nORB small-footprint real-time object request broker; the CyberMech platform (collaborative with Purdue University) for parallel Real-Time Hybrid Simulation; and the RT-Xen real-time virtualization research platform, from which the RTDS scheduler was transitioned into the Xen software distribution.

Yevgeniy Vorobeychik joined Washington University in St. Louis in 2018. He was an assistant professor of computer science and biomedical informatics at Vanderbilt University from 2013 until 2018, and a principal research scientist at Sandia National Laboratories from 2010 until 2013. Between 2008 and 2010 he was a postdoctoral research associate at the University of Pennsylvania Computer and Information Science department. He received a PhD and MSE in Computer Science and Engineering from the University of Michigan and a BS degree in Computer Engineering from Northwestern University. He received an NSF CAREER award in 2017 and was invited to give an IJCAI-16 early career spotlight talk. He was nominated for the 2008 ACM Doctoral Dissertation Award and received honorable mention for the 2008 IFAAMAS Distinguished Dissertation Award.

Dr. Xuan 'Silvia' Zhang is an Assistant Professor in the Preston M. Green Department of Electrical and Systems Engineering at Washington University in St. Louis. Before joining Washington University, she was a Postdoctoral Fellow in Computer Science at Harvard University. She received her B. Eng. degree in Electrical Engineering in 2006 from Tsinghua University in China, and her MS and PhD degree in Electrical and Computer Engineering from Cornell University in 2009 and 2012 respectively. She works across the fields of VLSI, computer architecture, and cyber physical systems and her research interests include adaptive power and resource management for autonomous systems, hardware/software co-design for machine learning and artificial intelligence, and efficient computation and security primitives in analog and mixed-signal domain. Dr. Zhang is the recipient of DATE Best Paper Award in 2019 and ISLPED Design Contest Award in 2013, and her work has also been nominated for Best Paper Award at DATE 2019 and DAC 2017.