Simulation to Scaled City: Zero-Shot Policy Transfer for Traffic Control via Autonomous Vehicles

Kathy Jang
University of California, Berkeley
kathyjang@berkeley.edu

Eugene Vinitsky
University of California, Berkeley
evinitsky@berkeley.edu

Behdad Chalaki
University of Delaware
bchalaki@udel.edu

Ben Remer
University of Delaware
bremer@udel.edu

Logan Beaver
University of Delaware
lebeaver@udel.edu

Andreas A. Malikopoulos
University of Delaware
andreas@udel.edu

Alexandre Bayen
Univesity of California, Berkeley
bayen@berkeley.edu

ABSTRACT

Using deep reinforcement learning, we successfully train a set of two autonomous vehicles to lead a fleet of vehicles onto a roundabout and then transfer this policy from simulation to a scaled city without fine-tuning. We use *Flow*, a library for deep reinforcement learning in microsimulators, to train two policies, (1) a policy with noise injected into the state and action space and (2) a policy without any injected noise. In simulation, the autonomous vehicles learn an emergent metering behavior for both policies which allows smooth merging. We then directly transfer this policy without any tuning to the *University of Delaware's Scaled Smart City (UDSSC)*, a 1:25 scale testbed for connected and automated vehicles. We characterize the performance of the transferred policy based on how thoroughly the ramp metering behavior is captured in UDSSC. We show that the noise-free policy results in severe slowdowns and only, occasionally, it exhibits acceptable metering behavior. On the other hand, the noise-injected policy consistently performs an acceptable metering behavior, implying that the noise eventually aids with the zero-shot policy transfer. Finally, the transferred, noise-injected policy leads to a 5% reduction of average travel time and a reduction of 22% in maximum travel time in the UDSSC. Videos of the proposed self-learning controllers can be found at https://sites.google.com/view/iccps-policy-transfer.

CCS CONCEPTS

• **Computing methodologies** → **Computational control theory**; **Reinforcement learning**; **Machine learning algorithms**; • **Computer systems organization** → *Robotic control*; *Robotic autonomy*; *Sensors and actuators*; *Dependable and fault-tolerant systems and networks*;

KEYWORDS

Cyber-physical systems, Deep learning, Reinforcement learning, Control theory, Autonomous vehicles, Policy Transfer

ACM Reference Format:
Kathy Jang, Eugene Vinitsky, Behdad Chalaki, Ben Remer, Logan Beaver, Andreas A. Malikopoulos, and Alexandre Bayen. 2019. Simulation to Scaled City: Zero-Shot Policy Transfer for Traffic Control via Autonomous Vehicles. In *10th ACM/IEEE International Conference on Cyber-Physical Systems (with CPS-IoT Week 2019) (ICCPS '19), April 16–18, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3302509.3313784

ICCPS '19, April 16–18, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6285-6/19/04…$15.00
https://doi.org/10.1145/3302509.3313784

1 INTRODUCTION

Control of mixed-autonomy traffic: Transportation is a major source of US energy consumption and greenhouse gas emissions, accounting for 28% and 26% respectively. According to the bureau of transportation statistics, total road miles traveled is continuously increasing, growing at 2 to 3% per year between 2010 and 2014 while over the same period the total road length of the US transportation network remained unchanged. The increased road usage is coupled with an increase in congestion. Overall congestion delay in 2014 was 6.9 billion hours, an increase of 33% since 2000; the problem is even worse in metropolitan areas where travelers needed to allocate an additional 150% more travel time during peak periods to arrive on time. The congestion also has significant economic cost, totaling 160 billion dollars in 2014 [5]. Depending on their usage, automated vehicles have the potential to alleviate system level metrics such as *congestion*, *accident rates*, and *greenhouse gas emissions* through a combination of intelligent routing, smoother driving behavior, and faster reaction time [27].

Partially automated systems are predicted to increasingly populate roadways between 2020 and 2025 but will primarily be usable in high driving or high speed operations in light traffic. Hazard detection technology is not expected to be mature enough for full automation in the presence of general vehicles and pedestrians (i.e. heterogeneous fleets, manned/unmanned, bicycles, pedestrians, mixed use road-space etc.) until at least 2030. It takes 20 years for a vehicle fleet to turn over sufficiently which makes it likely that vehicles will be partially manned at least until 2050 [20].

Recently, the steady increase in usage of cruise control systems on the roadway offers an opportunity to study the optimization

of traffic in the framework of *mixed-autonomy traffic*: traffic that is partially automated but mostly still consists of human driven vehicles. However, the control problems posed in this framework are notoriously difficult to solve. Traffic problems, which often exhibit features such as time-delay, non-linear dynamics, and hybrid behavior, are challenging for classical control approaches, as microscopic traffic models are high complexity: discrete events (lane changes, traffic light switches), continuous states (position, speed, acceleration), and non-linear driving models. These complexities make analytical solutions often intractable. The variety and non-linearity of traffic often leads to difficult trade-offs between the fidelity of the dynamics model and tractability of the approach.

Classical control approaches: Classical control approaches have successfully solved situations in which the complexity of the problem can be reduced without throwing away key aspects of the dynamics. For example, there is a variety of analytical work on control of autonomous intersections with simple geometries. For mixed-autonomy problems, there have been significant classical controls based results for simple scenarios like vehicles on a ring [4] or a single lane of traffic whose stability can be characterized [14, 23]. A thorough literature review on coordinating autonomous vehicles in intersections, merging roadways, and roundabouts can be found in [17]. The classical control approaches described in this review can be broken down into reservation methods, scheduling, optimization with safety constraints, and safety maximization. Other approaches discussed in the review involve applications of queuing theory, game theory, and mechanism design.

However, as the complexity of the problem statement increases, classical techniques become increasingly difficult to apply. Shifting focus from simple scenarios to, for example, hybrid systems with coexisting continuous and discrete controllers, explicit guarantees for hand-designed controllers can become harder to find. Ultimately, when the complexity of the problem becomes too high, optimization-based approaches have been shown to be a successful approach in a wide variety of domains from robotics [9] to control of transportation infrastructure [11].

Deep reinforcement learning: *Deep reinforcement learning* (deep RL) has recently emerged as an effective technique for control in high dimensional, complex CPS systems. Deep RL has shown promise for the control of complex, unstructured problems as varied as robotic skills learning [7], playing games such as Go [21], and traffic light ramp metering [1]. Of particular relevance to this work, deep RL has been successful in training a single autonomous vehicle to optimize traffic flow in the presence of human drivers [28].

One key distinction in RL is whether the algorithm is model-free or model-based, referring to whether the algorithm is able to query a dynamics model in the computation of the control or the policy update. Model-free RL tends to outperform model-based RL if given sufficient optimization time, but requires longer training times. Thus, model-free techniques are most effective when samples can be cheaply and rapidly generated. This often means that model-free RL works best in simulated settings where a simulation step can be made faster than real-time and simulation can be distributed across multiple CPUs or GPUs. A long-standing goal is to be able to train a controller in simulation, where model-free techniques can be used, and then use the trained controller to control the actual system.

Policy Transfer: Transfer of a controller from a training domain to a new domain is referred to as *policy transfer*. The case where the policy is directly transferred without any fine-tuning is referred to as *zero-shot policy transfer*. Zero-shot policy transfer is a difficult problem in RL, as the true dynamics of the system may be quite different from the simulated dynamics, an issue referred to as *model mismatch*. Techniques used to overcome this include adversarial training, in which the policy is trained in the presence of an adversary that can modify the dynamics and controller outputs and the policy must subsequently become robust to perturbations [16]. Other techniques to overcome *model mismatch* include re-learning a portion of the controller[18], adding noise to the dynamics model [15], and learning a model of the true dynamics that can be used to correctly execute the desired trajectory of the simulation-trained controller [3]. Efforts to overcome the reality gap have been explored in vision-based reinforcement learning [13] and in single AV systems [29].

Other challenges with policy transfer include *domain mismatch*, where the true environment contains states that are unobserved or different from simulation. For example, an autonomous vehicle might see a car color that is unobserved in its simulations and subsequently react incorrectly. Essentially, the controller overfits to its observed states and does not generalize. Domain mismatch can also occur as a result of imperfect sensing or discrepancies between the simulation and deployment environment. While in simulation it is possible to obtain perfect observations, this is not always the case in the real world. Small differences between domains can lead to drastic differences in output. For example, a slight geometric difference between simulation and real world could result in a vehicle being registered as being on one road segment, when it is on another. This could affect the control scheme in a number of ways, such as a premature traffic light phase change. Techniques used to tackle this problem include domain randomization [24], in which noise is injected into the state space to enforce robustness with respect to unobserved states.

Contributions and organization of the article: In this work we use deep RL to train two autonomous vehicles to learn a classic form of control: ramp metering, in which traffic flow is regulated such that one flow of vehicles is slowed such that another flow can travel faster. While in the real world, ramp metering is controlled via metering lights, we demonstrate the same behavior using AVs instead of lights. Each RL vehicle interacts with sensors at each of the entrance ramps and is additionally able to acquire state information about vehicles on the roundabout, as well as state information about the other RL vehicle. By incorporating this additional sensor information, we attempt to learn a policy that can time the merges of the RL vehicles and their platoons to learn ramp metering behavior, which prevents energy-inefficient decelerations and accelerations. Being positioned at the front of a platoon of vehicles, each RL vehicle has the ability to control the behavior of the platoon of human-driven vehicles following it. The RL vehicle, also referred to in this paper as an autonomous vehicle (AV), is trained with the goal of minimizing the average delay of all the vehicles in simulation.

Next, we show how we overcome the RL to real world reality gap and demonstrate RL's real world relevance by transferring the controllers to the University of Delaware's Scaled Smart City (UDSSC),

a reduced-scale city whose dynamics, which include sensor delays, friction, and actuation, are likely closer to true vehicle dynamics. RL trained policies, which are learned in a simulation environment, can overfit to the dynamics and observed states of the simulator and can then fare poorly when transferred to the real world. The combination of model and domain mismatch contributes to this problem. We combine the ideas of domain randomization with adversarial perturbations to the dynamics and train a controller in the presence of noise in both its observations and actions. For reasons discussed in Sec. 4, we expect the addition of noise in both state and action to help account for both model and domain mismatch.

In this work we present the following results:

- The use of deep RL in simulation to learn an emergent metering policy.
- A demonstration that direct policy transfer to UDSSC leads to poor performance.
- A successful zero-shot policy transfer of the simulated policy to the UDSSC vehicles via injection of noise into both the state and action space.
- An analysis of the improvements that the autonomous vehicles bring to the congested roundabout.

The remainder of the article is organized as follows:

(1) Section 2 provides an introduction to deep RL and the algorithms used in this work.
(2) Section 3 describes the setup we use to learn control policies via RL, followed by the policy transfer process from simulation to the physical world.
(3) Section 4 discusses the results of our experiments and provides intuition for the effectiveness of the state and action noise.
(4) Section 5 summarizes our work and future directions.

2 BACKGROUND

2.1 Reinforcement Learning

In this section, we discuss the notation and briefly describe the key concepts used in RL. RL focuses on deriving optimal controllers for *Markov decision processes* (MDP) [2]. The system described in this article solves tasks which conform to the standard structure of a finite-horizon discounted MDP, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma, T)$. Here $\mathcal{S}$ is a set of states and $\mathcal{A}$ is a set of actions where both sets can be finite or infinite. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the transition probability distribution describing the probability of moving from one state s to another state s' given action a, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the probability distribution over start states, $\gamma \in (0, 1]$ is the discount factor, and T is the horizon. For partially observable tasks, which conform to the structure of a *partially observable Markov decision process* (POMDP), two more components are required, namely Ω, a set of observations of the hidden states, and $O : \mathcal{S} \times \Omega \to \mathbb{R}_{\geq 0}$, the observation probability distribution.

RL studies the problem of how an agent can learn to take actions in its environment to maximize its expected cumulative discounted reward: specifically it tries to optimize $R = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$ where r_t is the reward at time t. The goal is to use the observed data from the MDP to optimize a *policy* $\Pi : \mathcal{S} \to \mathcal{A}$, mapping states to actions, that maximizes R. This policy can be viewed as the controller for the system, however, we stick to the convention of RL literature and refer to it as a policy. It is increasingly common to parameterize the policy via a neural net. We will denote the parameters of this policy, which are the weights of the neural network, by θ and the policy by π_θ. A neural net consists of a stacked set of affine linear transforms and non-linearities that the input is alternately passed through. The presence of multiple stacked layers is the origin of the term "deep"-RL.

2.2 Policy Gradient Methods

Policy gradient methods use Monte Carlo estimation to compute an estimate of the gradient of the expected discounted reward $\nabla_\theta R = \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$ where θ are the parameters of the policy π_θ. We perform repeated *rollouts*, in which the policy is used to generate the actions at each time step. At the end of the rollout, we have accumulated a state, action, reward trajectory $\tau = (s_0, a_0, r_0, \ldots, s_T)$. Policy gradient methods take in a set of these trajectories and use them to compute an estimate of the gradient $\nabla_\theta R$ which can be used in any gradient ascent-type method.

The particular policy gradient method used in this paper is *Trust Region Policy Optimization* (TRPO) [19]. TRPO is a monotonic policy improvement algorithm, whose update step provides guarantees of an increase in the expected total reward. However, the exact expression for the policy update leads to excessively small steps so implementations of TRPO take larger steps by using a trust region. In this case, the trust region is a bound on the KL divergence between the old policy and the policy update. While not a true distance measure, a small KL divergence between the two policies suggests that the policies do not act too differently over the observed set of states, preventing the policy update step from sharply shifting the policy behavior.

2.3 Car Following Models

For our model of the driving dynamics, we used the *Intelligent Driver Model* [25] (IDM) that is built into the traffic microsimulator SUMO [8]. IDM is a microscopic car-following model commonly used to model realistic driver behavior. Using this model, the acceleration for vehicle α is determined by its bumper-to-bumper *headway* s_α (distance to preceding vehicle), the vehicle's own velocity v_α, and relative velocity Δv_α, via the following equation:

$$a_{\text{IDM}} = \frac{dv_\alpha}{dt} = a\left[1 - \left(\frac{v_\alpha}{v_0}\right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s_\alpha}\right)^2\right] \tag{1}$$

where s^* is the desired headway of the vehicle, denoted by:

$$s^*(v_\alpha, \Delta v_\alpha) = s_0 + \max\left(0, v_\alpha T + \frac{v_\alpha \Delta v_\alpha}{2\sqrt{ab}}\right) \tag{2}$$

where $s_0, v_0, T, \delta, a, b$ are given parameters. Typical values for these parameters can be found in [25]; the values used in our simulations are given in Sec. 3.1.1. To better model the natural variability in driving behavior, we induce stochasticity in the desired driving speed v_0. For a given vehicle, the value of v_0 is sampled from a Gaussian whose mean is the speed limit of the lane and whose standard deviation is 20% of the speed limit.

Car following models are not inherently collision-free, we supplement them with a safe following rule: a vehicle is not allowed to

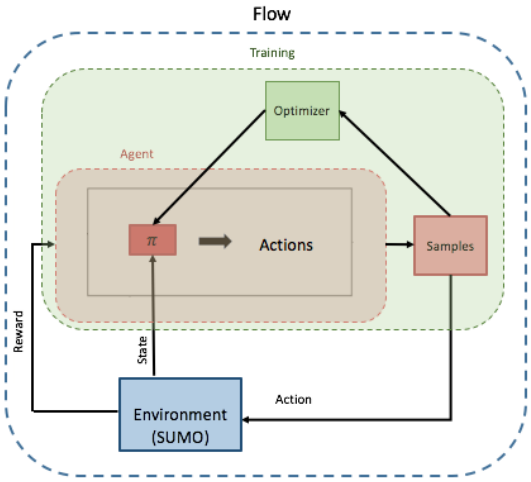

Figure 1: Diagram of the iterative process in *Flow*. Portions in red correspond to the controller and rollout process, green to the training process, and blue to traffic simulation.

take on velocity values that might lead to a crash if its lead vehicle starts braking at maximum deceleration. However, due to some uncertainty in merging behavior, there are still rare crashes that can occur in the system.

2.4 Flow

We run our experiments in *Flow* [28], a library that provides an interface between a traffic microsimulator, SUMO [8], and two RL libraries, rllab [6] and RLlib [12], which are centralized and distributed RL libraries respectively. *Flow* enables users to create new traffic networks via a Python interface, introduce autonomous controllers into the networks, and then train the controllers in a distributed system on the cloud via AWS EC2. To make it easier to reproduce our experiments or try to improve on our benchmarks, the code for *Flow*, scripts for running our experiments, and tutorials can be found at https://github.com/flow-project/flow.

Fig. 1 describes the process of training the policy in *Flow*. The controller, here represented by policy π_θ, receives a state and reward from the environment and uses the state to compute an action. The action is taken in by the traffic microsimulator, which outputs the next state and a reward. The (state, next state, action, reward) tuple are stored as a sample to be used in the optimization step. After accumulating enough samples, the states, actions, and rewards are passed to the optimizer to compute a new policy.

2.5 University of Delaware's Scaled Smart City (UDSSC)

The *University of Delaware's Scaled Smart City (UDSSC)* was used to validate the performance of the RL control system. UDSSC is a testbed (1:25 scale) that can help prove concepts beyond the simulation level and can replicate real-world traffic scenarios in a small and controlled environment. UDSSC uses a VICON camera system to track the position of each vehicle with sub-millimeter accuracy, which is used both for control and data collection. The

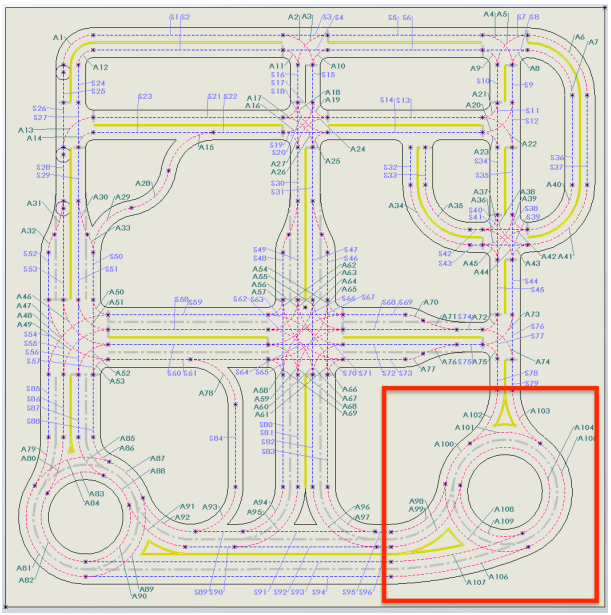

Figure 2: Diagram of the UDSSC road map, with the experimental zone highlighted in red.

controller for each vehicle is offloaded to a mainframe computer and runs on an independent thread which is continuously fed data from the VICON system. Each controller uses the global VICON data to generate a speed reference for the vehicles allowing for precise independent closed-loop feedback control. A detailed description of UDSSC can be found in [22]. To validate the effectiveness of the proposed RL approach in a physical environment, the southeast roundabout of the UDSSC was used (Fig. 2).

3 EXPERIMENTAL DEPLOYMENT

3.1 Experimental setup

3.1.1 Simulation Details. To derive the RL policy, we developed a model of the roundabout highlighted in red in Fig. 2 in SUMO. The training of the model, shown in Fig. 3, included a single-lane roundabout with entry points at the northern and western ends. Throughout this paper we will refer to vehicles entering from the western end as the *western platoon* and the north entrance as the *northern platoon*. The entry points of the model are angled slightly different as can be seen in Figs. 2 and 3.

The human-controlled vehicles operate using SUMO's built-in IDM controller, with several modified parameters. In these experiments, the vehicles operating with the IDM controller are run with $T = 1$, $a = 1$, $b = 1.5$, $\delta = 4$, $s_0 = 2$, $v_0 = 30$, and noise = 0.1, where T is a safe time headway, a is a comfortable acceleration in m/s^2, b is a comfortable deceleration, δ is an acceleration exponent, $s0$ is the linear jam distance, v_0 is a desired driving velocity, and noise is the standard deviation of a zero-mean normal perturbation to the acceleration or deceleration. Details of the physical interpretation of these parameters can be found in [25]. Environment parameters in simulation were set to match the physical constraints of UDSSC.

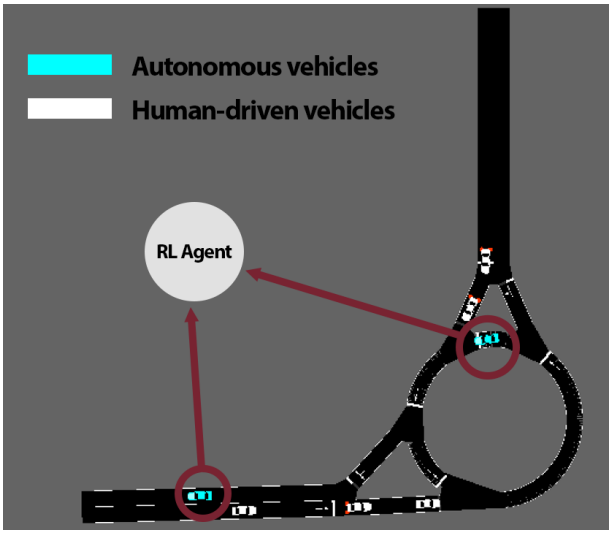

Figure 3: SUMO-generated network in UDSSC's roundabout. The blue vehicles are the AVs; they are both controlled by the RL policy. Videos of this policy in simulation are available at https://sites.google.com/view/iccps-policy-transfer.

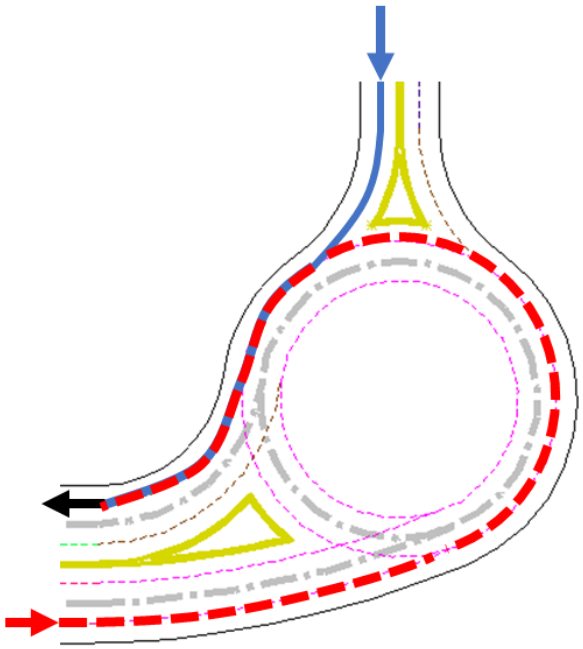

Figure 4: Visualization of the path taken on the UDSSC roundabout. The red route enters going east and exits going west; the blue route enters north and exits via the same west entrance as the red route.

These include: a maximum acceleration of $1\frac{\text{m}}{\text{s}^2}$, a maximum deceleration of $-1\frac{\text{m}}{\text{s}^2}$, and a maximum velocity of $15\frac{\text{m}}{\text{s}}$. The timestep of the system is set to 1.0 seconds.

Simulations on this scenario in the roundabout were executed across a range of different settings in terms of volume and stochasticity of inflows. In the RL policy implemented in UDSSC and discussed in 3.1.2, vehicles are introduced to the system via deterministic inflows from the northern and western ends of the roundabout using two routes: (1) the northern platoon enters the system from the northern inflow, merges into the roundabout, and exits through the western outflow and (2) the western platoon enters the system from the western inflow, U-turns through the roundabout, and exits through the western outflow. The western platoon consists of four vehicles total: three vehicles controlled with the IDM controller led by a vehicle running with the RL policy. The northern platoon consists of of three vehicles total: two vehicles controlled with the IDM controller led by a vehicle running with the RL policy. New platoons enter the system every 1.2 minutes, the rate of which is significantly sped up in simulation. These inflow settings are designed to showcase the scenario where routes clash (Fig. 3).

3.1.2 UDSSC. Each vehicle in UDSSC uses a saturated IDM controller to (1) avoid negative speeds, (2) ensure that the rear-end collision constraints do not become active, and (3) maintain the behavior of Eq. (1). Both the IDM and RL controllers provide a desired acceleration for the vehicles, which is numerically integrated to calculate each vehicle's reference speed.

Merging at the northern entrance of the roundabout is achieved by an appropriate yielding function. Using this function, the car entering the roundabout proceeds only if no other vehicle is on the roundabout at a distance from which a potential lateral collision may occur. Otherwise, the vehicle stops at the entry of the roundabout waiting to find a safe space to proceed.

To match the SUMO training environment, only one vehicle per path was allowed to use the RL policy. The paths taken by each vehicle are shown in Fig. 4. For each path, the first vehicle to enter the experimental zone was controlled by the RL policy; every subsequent vehicle runs with the saturated IDM controller. Once an active vehicle running with the RL policy exits the experimental zone, it reverts back to the IDM controller.

In the experiments, the vehicles operated in a predefined deterministic order, as described in 3.1.1. Four vehicles were placed just outside the experimental zone on the western loop, and three vehicles were placed in the same fashion near the northern entrance. Each vehicle platoon was led by a vehicle running with the RL policy, except in the baseline case where all vehicles used the saturated IDM controller. The experiment was executed with three variations: (1) the baseline case with all vehicles running with the IDM controller, (2) the case with a leader vehicle running with an RL policy trained in SUMO, and (3) the case where the leader vehicles running with an RL policy were trained in simulation with noise injected into their observations and accelerations.

3.2 Reinforcement Learning Structure

3.2.1 Action space. We parametrize the controller as a neural net mapping the observations to a mean and diagonal covariance matrix of a Gaussian. The actions are sampled from the Gaussian; this is a standard controller parametrization [10]. The actions are a

two-dimensional vector of accelerations in which the first element corresponds to vehicles on the north route and the second element to the west route. Because the dimension of the action vector is fixed, there can only ever be 1 AV from the northern entry and 1 AV from the western entry. Two queues, one for either entryway, maintain a list of the RL-capable vehicles that are currently in the system. It should be noted that the inflow rates are chosen such that the trained policy never contains more than a queue of length 2. Platoons are given ample time to enter and exit the system before the next platoon arrives. This queue mechanism is designed to support the earlier stages of training, when RL vehicles are learning how to drive, which can result in multiple sets of platoons and thus more than 2 RL vehicles being in the system at the same time. Control is given to vehicles at the front of both queues. When a vehicle completes its route and exits the experimental zone, its ID is popped from the queue. All other RL-capable vehicles are passed IDM actions until they reach the front of the queue. If there are fewer than two AVs in the system, the extra actions are simply unused.

The dynamics model of the autonomous vehicles are given by the IDM described in sec. 2.3 subject to a minimum and maximum speed i.e.

$$v_j^{\text{IDM}}(t+\Delta t) = \max\left(\min\left(v_{AV}(t) + a_{IDM}\Delta t, v_j^{\max}(t)\right), 0\right) \quad (3)$$

where $v_j^{AV}(t)$ is the velocity of autonomous vehicle j at time t, a_{IDM} is the acceleration given by an IDM controller, Δt is the time-step, and $v_j^{\max}(t)$ is the maximum speed set by the city j. For the AVs, the acceleration a_t is straightforwardly added to the velocity via a first-order Euler integration step

$$v_j^{\text{AV}}(t+\Delta t) = \max\left(\min\left(v_j^{\text{AV}}(t) + a_t\Delta t, v_j^{\max}(t)\right), 0\right) \quad (4)$$

3.2.2 Observation space. For the purposes of keeping in mind physical sensing constraints, the state space of the MDP is partially observable. It is normalized to ±1 and includes the following:

- The positions of the AVs.
- The velocities of the AVs.
- The distances from the roundabout of the 6 closest vehicles to the roundabout for both roundabout entryways.
- The velocities of the 6 closest vehicles to the roundabout for both roundabout entryways.
- Tailway and headway (i.e. distances to the leading and following vehicles) of vehicles from both AVs.
- Length of the number of vehicles waiting to enter the roundabout for both roundabout entryways.
- The distances and velocities of all vehicles in the roundabout.

This state space was designed with real-world implementation in mind, and could conceivably be implemented on existing roadways equipped with loop detectors, sensing tubes, and vehicle-to-vehicle communication between the AVs. For a sufficiently small roundabout, it is possible that an AV equipped with enough cameras could identify the relevant positions and velocities of roundabout vehicles. Similarly, the queue lengths can be accomplished with loop detectors, and the local information of the AVs (its own position and velocity, as well as the position and velocity of its leader and follower) are already necessarily implemented in distance-keeping cruise control systems.

3.2.3 Action and State Noise. The action and state spaces are where we introduce noise with the purpose of training a more generalizable policy that is more resistant to the difficulties of cross-domain transfer. We train the policies in two scenarios, a scenario where both the action and state space are perturbed with noise and a scenario with no noise. This former setting corresponds to a type of *domain randomization*. In the noisy case, we draw unique perturbations for each element of the action and state space from a Gaussian distribution with zero mean and a standard deviation of 0.1. In the action space, which is composed of just accelerations, this corresponds to a standard deviation of $0.1\frac{\text{m}}{\text{s}^2}$. In the state space, which is normalized to 1, this corresponds to a standard deviation of 1.5 m/s for velocity-based measures. The real-life deviations of each distance-based state space element are described here: AV positions, and the tailways and headways of the AVs, deviate the most at 44.3 m, the large uncertainty of which results in a policy that plays it safe. The distance from the northern and western entryways respectively deviate by 7.43m and 8.66 m. The length of the number of vehicles waiting to enter the roundabout from the northern and western entryway respectively deviate by 1.6 and 1.9 vehicles.

These perturbations are added to each element of the action and state space. The elements of the action space are clipped to the maximum acceleration and deceleration of ±1, while the elements of the state space are clipped to ±1 to maintain normalized boundaries. Noisy action and state spaces introduce uncertainty to the training process. The trained policy must still be effective even in the presence of uncertainty in its state as well as uncertainty that its requested actions will be faithfully implemented.

3.2.4 Reward function. For our reward function we use a combination of the L2-norm of the velocity of all vehicles in the system and penalties discouraging standstills or low velocity travel.

$$r_t = \frac{\max\left(v_{\max}\sqrt{n} - \sqrt{\sum_{i=1}^{n}(v_{i,t} - v_{\max})^2}, 0\right)}{v_{\max}\sqrt{n}} - 1.5 \cdot \text{pen}_\text{s} - \text{pen}_\text{p} \quad (5)$$

where n is the number of all vehicles in the system, $v_{\max}$ is the maximum velocity of $15\frac{\text{m}}{\text{s}}$, $v_{i,t}$ is the velocity that vehicle v_i is travelling at at time t. The first term incentives vehicles to travel near speed $v_{\max}$ but also encourages the system to prefer a mixture of low and high velocities versus a mixture of mostly equal velocities. The preference for low and high velocities is intended to induce a platooning behavior. RL algorithms are sensitive to the scale of the reward functions; to remove this effect the reward is normalized by $v_{\max}\sqrt{n}$ so that the maximum reward of a time-step is 1.

This reward function also introduces 2 penalty functions, pen_s and pen_p. pen_s returns the number of vehicles that are traveling at a velocity of 0, and pen_p is the number of vehicles that are traveling below a velocity of 0.3 m/s. They are defined as:

$$\text{pen}_\text{s} = \sum_{i=1}^{n} g(i) \text{ where } g(x) = \begin{cases} 0, & v_x \neq 0, \\ 1, & v_x = 0. \end{cases} \quad (6)$$

$$\text{pen}_p = \sum_{i=1}^{n} h(i) \text{ where } h(x) = \begin{cases} 0, & v_x \geq 0.3, \\ 1, & v_x \leq 0.3 \end{cases} \quad (7)$$

These penalty functions are added to discourage the autonomous vehicle from fully stopping or adopting near-zero speeds. In the absence of these rewards, the RL policy learns to game the simulator by blocking vehicles from entering the simulator on one of the routes, which allows for extremely high velocities on the other route. This occurs because velocities of vehicles that have not yet emerged from an inflow do not register, so no penalties are incurred when the AV blocks further vehicles from entering the inflow.

3.3 Algorithm/simulation details

We ran the RL experiments with a discount factor of .999, a trust-region size of .01, a batch size of 20000, a horizon of 500 seconds, and trained over 100 iterations. The controller is a neural network, a *Gaussian multi-layer perceptron* (MLP), with hidden sizes of (100, 50, 25) and a tanh non-linearity. The choice of neural network non-linearities, size, and type were picked based on traffic controllers developed in [26]. The states are normalized so that they are between 0 and 1 by dividing each states by its maximum possible value. The actions are clipped to be between −1 and 1. Both normalization and clipping occur after the noise is added to the system so that the bounds are properly respected.

3.4 Code reproducibility

In line with open, reproducible science, the following codebases are needed to reproduce the results of our work. *Flow* can be found at https://github.com/flow-project/flow. The version of *rllab* used for the RL algorithms is available at https://github.com/cathywu/rllab-multiagent at commit number **4b5758f**. *SUMO* can be found at https://github.com/eclipse/sumo at commit number **1d4338ab80**.

3.5 Policy Transfer

The RL policy learned through Flow was encoded as the weights of a neural network. These weights were extracted from a serialized file and accessed via a Python function which maps inputs and outputs identical to those used in training. Separating these weights from rllab enables an interface for state space information from the UDSSC to be piped straight into the Python function, returning the accelerations to be used on the UDSSC vehicles. The Python function behaves as a control module within the UDSSC, replacing the IDM control module in vehicles operating under the RL policy.

The inputs to the RL neural network were captured by the VICON system and mainframe. The global 2D positions of each vehicle were captured at each time step. These positions were numerically derived to get each vehicle's speed and were compared to the physical bounds on the roadways to get the number of vehicles in each queue at the entry points. Finally, the 2D positions were mapped into the 1D absolute coordinate frame used during training. This array was passed into the RL control module as the inputs of the neural network.

3.6 Results

In this section we present our results from a) training vehicular control policies via RL in simulation, and b) transferring the successful policy to UDSSC. Extended work and videos of the policies in action are available at https://sites.google.com/view/iccps-policy-transfer.

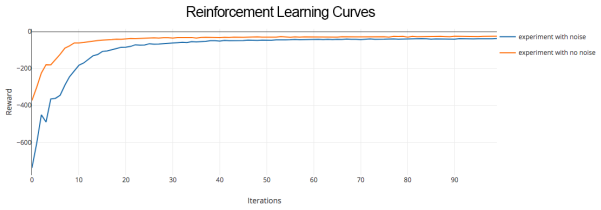

Figure 5: Convergence of the RL reward curve of an experiment with noised IDM, RL accelerations, and noisy state space

3.6.1 Simulation results. Fig. 5 depicts the reward curve. The noise-injected RL policy takes longer to train than the noise-free policy and fares much worse during initial training, but converges to an almost identical final reward. In the simulations, videos of which are on the website, a ramp metering behavior emerges in which the incoming western vehicle learns to slow down to allow the vehicles on the north ramp to smoothly merge.

This ramp metering behavior can also be seen in the space-time diagrams in Fig. 6, which portrays the vehicle trajectories and velocities of each vehicle in the system. Western vehicles are depicted in blue and northern in red. Due to the overlapping routes, visible in Fig. 4, it was necessary to put a kink in the diagram for purposes of clarity; the kink is at the point where the northern and western routes meet. As can be seen in the middle figure, in the baseline case the two routes conflict as the northern vehicles aggressively merge onto the ramp and cut off the western platoon. Once the RL policy controls the autonomous vehicles, it slows down the western platoon so that no overlap occurs and the merge conflict is removed.

3.6.2 Transfer to UDSSC. The RL policies were tested under three cases in UDSSC: (1) the baseline case with only vehicles running with the IDM controller, (2) the case with a leader vehicle running with the RL policy trained in sumo without additional noise, and (3) the case where the leader vehicles running with the RL policy were trained with noise actively injected into their observations and accelerations. The outcomes of these trials are presented in Table 1.

During the congestion experiment, the third case, in which noise was actively injected into the action and state space during training, successfully exhibits the expected behavior it demonstrated in simulation. In this RL controlled case, the transfer consistently showed successful ramp metering: the western platoon adopted a lower speed than the baseline IDM controller, as can be seen in the lower velocity of the RL vehicle in Fig. 7. This allowed the northern queue to merge before the western platoon arrived, increasing the overall throughput of the roundabout. This is closer to a socially optimal behavior, leading to a lower average travel time than the greedy behavior shown in the baseline scenario. No unexpected or dangerous driving behavior occurred.

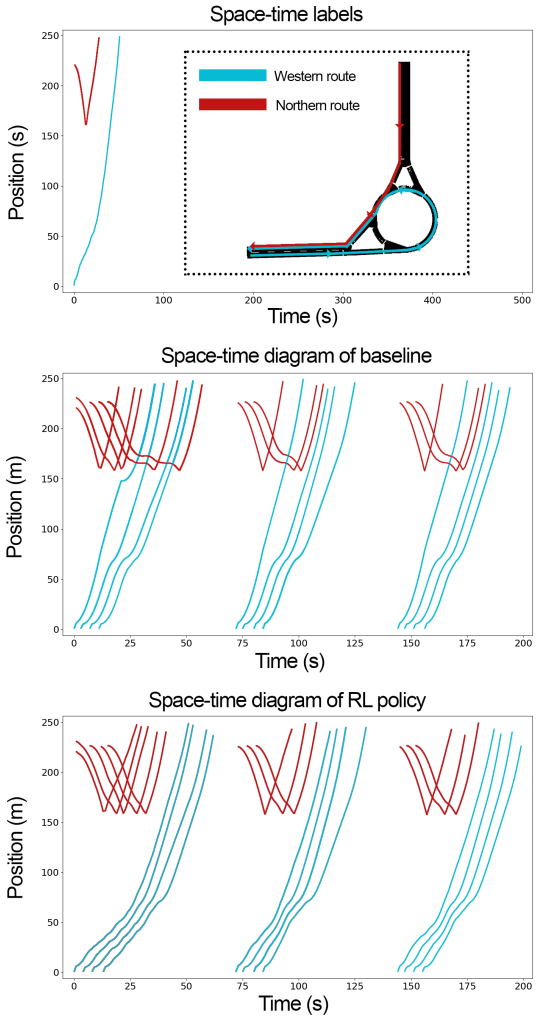

Figure 6: Space-time diagrams of the simulated baseline and RL policy. Each line corresponds to a vehicle in the system. Top: a guide to the color-scheme of the space time diagrams. The northern route is in red, the western route in blue. Middle: illustrates the overlap between the merging northern platoon and the western platoon. Bottom: The RL policy, depicted at the bottom successfully removes this overlap. Videos of this policy in simulation are available at https://sites.google.com/view/iccps-policy-transfer.

This is in comparison with the second case, a policy trained on noiseless observations. In the second case, undesirable and unexpected deployment behavior suggests problems with the transfer process. Collisions occurred, sometimes leading to pile-ups, and platooning would frequently be timed incorrectly, such that, for example, only part of the Western platoon makes it through the roundabout before the Northern platoon cuts the Western platoon off. This indicates that the noise-injected policy is robust to transfer and resistant to domain and model mismatch.

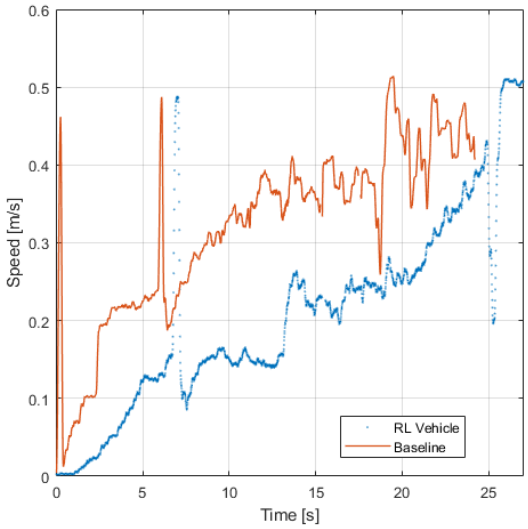

Figure 7: Comparison of the first vehicle on the southern loop for the baseline (IDM) and RL experiments. The RL vehicle starts off slower but eventually accelerates sharply once the northern platoon has passed.

Figure 8: Experiment with two platoons being led by RL vehicles (blue, circled).

Furthermore, the platoons led by RL vehicles trained with injected noise outperformed the baseline and noise-free cases. The results of these experiments, averaged over three trials with the RL vehicles, are presented in Table 1. This improvement was the outcome of a metering behavior learned by the western RL platoon leader. In the baseline case, the north and western platoons meet and lead to a merge conflict that slows the incoming western vehicles down. This sudden decrease in speed can be seen in the drop in velocity at 20 seconds of the baseline in Fig. 7.

Fig. 8 shows the experiment in progress, with the blue (circled) vehicles being RL vehicles trained under noisy conditions. The RL

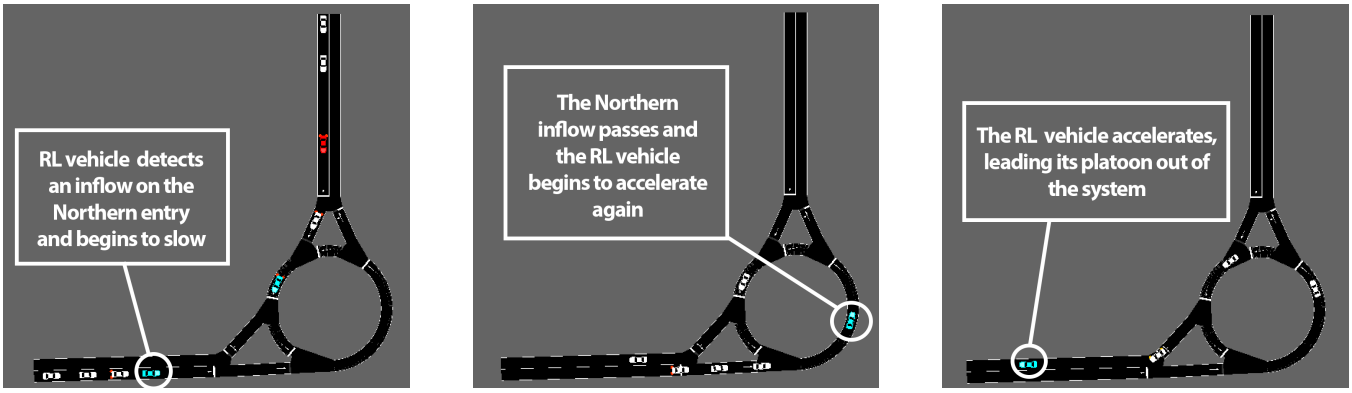

Figure 9: **RL-controlled vehicle demonstrating smoothing behavior in this series of images. First: RL vehicle slows down in anticipation of a sufficiently short inflow from the north. Second: The northern inflow passes through the roundabout at high velocity. Fourth:The RL vehicle accelerates and leads its platoon away from the roundabout. Videos of this policy in simulation are available at** https://sites.google.com/view/iccps-policy-transfer.

	Avg. Vel.[m/s]	Avg. Time[s]	Max Time[s]
Baseline	0.26	15.71	23.99
RL	0.22	15.68	20.62
RL with Noise	0.23	14.81	18.68

Table 1: Results for the congestion experiment, the average and maximum times are averaged between three RL trials and a single baseline trial.

vehicle entering from the western (lower) entrance has performed it's metering behavior, allowing vehicles from the northern (upper) queue to pass into the roundabout before the western RL vehicle speeds up again. Videos of the emergent behavior can be found on the website.

4 DISCUSSION

For the simulated environment, the choice of reward function, specifically, using the L2-norm rather than the L1-norm, encourages a more stable, less sparse solution. This makes evaluating the success of policy transfer more straightforward. Table 1 reports the results of the UDSSC experiments on three metrics:

- The average velocity of the vehicles in the system
- The average time spent in the system
- The maximum time that any vehicle spent in the system

Note, the system is defined as the entire area of the experiments, including the entrances to the roundabouts. The layer of uncertainty in the noise-injected policy aided with overcoming the domain and model mismatch between the simulation system and the UDSSC system. Thus, the noised policy was able to successfully transfer from simulation to UDSSC and also improved the average travel time by 5% and the maximum travel time by 22%. The noise-free policy did not improve on the average travel time and only improved the maximum travel time by 14%. Although we did not perform an ablation study to check whether both state space noise and action space noise were necessary, this does confirm that the randomization improved the policy transfer process.

The UDSSC consistently reproduced the moderate metering behavior for the noise-injected policy, but did not do so for the policy that was trained without noise. As can be seen in the videos, the noise-free policy was not consistent and would only irregularly reproduce the desired behavior, or meter dramatically to the point that average travel time increased. Overall the noised policy significantly outperformed the noise-free version.

However, we caution that in our testing of the policy transfer process on the UDSSC, we performed a relatively limited test of the effectiveness of the policy. The vehicles were all lined up outside the system and then let loose; thus, the tests were mostly deterministic. Any randomness in the tests would be due solely to randomness in the dynamics of the UDSSC vehicles and stochasticity in the transferred policy. In training, the acceleration of IDM vehicles are noised and can account for some stochasticity in the initial distribution of vehicles on the UDSSC. However, the trained policy was not directly given this inflow distribution at train time, so this does correspond to a separation between train and test sets.

There may be several reasons why the noise and action injection may have allowed for a successful zero shot transfer. First, because the action is noisy, the learned policy will have to learn to account for *model mismatch* in its dynamics: it cannot assume that the model is exactly the double-integrator that is used in the simulator. Subsequently, when the policy is transferred to an environment with both delay, friction, and mass, the policy sees the mismatch as just another form of noise and accounts for it successfully. The addition of state noise helps with domain randomization; although the observed state distributions of the simulator and the scaled city may not initially overlap, the addition of noise expands the volume of observed state space in the simulator which may cause the two state spaces to overlap. Finally, the addition of noise forces the policy to learn to appropriately filter noise, which may help in the noisier scaled city environment.

5 CONCLUSIONS

In this paper, we demonstrated the real-world relevance of deep RL AV controllers for traffic control by overcoming the gap between simulation and the real world. Using RL policies, AVs in the UDSSC testbed successfully coordinated at a roundabout to ensure a smooth merge. We trained two policies, one in the presence of state and action space noise, and one without, demonstrating that the addition of noise led to a successful transfer of the emergent metering behavior, while the noise-free policy often over-metered, or failed, to meter at all. This implies that for non-vision based robotic systems with small action-dimension, small amounts of noise in state and action space may be sufficient for effective zero-shot policy transfer. As a side benefit, we also demonstrate that the emergent behavior leads to a reduction of 5% in average travel time and 22% max-travel time on the transferred network.

Ongoing work includes characterizing this result more extensively, evaluating the effectiveness of the efficiency of the policy against a wide range of vehicle spacing, platoon sizes, and inflow rates. In this context, there are still several questions we hope to address, as for example:

- Are both state and action space noise needed for effective policy transfer?
- What scale and type of noise is most helpful in making the policy transfer?
- Would selective domain randomization yield a less lossy, more robust transfer?
- Would adversarial noise lead to a more robust policy?
- Can we theoretically characterize the types of noise that lead to zero-shot policy transfer?

Finally, we plan to generalize this result to more complex roundabouts including many lanes, many entrances, and the ability of vehicles to change lanes. We also plan to evaluate this method of noise-injected transfer on a variety of more complex scenarios, such as intersections using stochastic inflows of vehicles.

ACKNOWLEDGMENTS

This research was funded in part by an NSF Graduate Research Fellowship and in part by the Delaware Energy Institute (DEI). AWS credits and funding were provided by an Amazon Machine Learning Research award.

The authors would also like to thank the Ray Zayas and Ishtiaque Mahbub for their contributions to enhancing the UDSSC database.

REFERENCES

[1] Francois Belletti, Daniel Haziza, Gabriel Gomes, and Alexandre M Bayen. 2017. Expert level control of ramp metering based on multi-task deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems* (2017).

[2] Richard Bellman. 1957. A Markovian decision process. *Journal of Mathematics and Mechanics* (1957), 679–684.

[3] Paul Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. 2016. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518* (2016).

[4] Shumo Cui, Benjamin Seibold, Raphael Stern, and Daniel B Work. 2017. Stabilizing traffic flow via a single autonomous vehicle: Possibilities and limitations. In *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 1336–1341.

[5] US DOT. 2016. National transportation statistics. *Bureau of Transportation Statistics, Washington, DC* (2016).

[6] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. 2016. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*. 1329–1338.

[7] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. 2017. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 3389–3396.

[8] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. 2012. Recent Development and Applications of SUMO - Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements* 5, 3&4 (December 2012), 128–138.

[9] Scott Kuindersma, Robin Deits, Maurice Fallon, Andrés Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. 2016. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous Robots* 40, 3 (2016), 429–455.

[10] Sergey Levine and Pieter Abbeel. 2014. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems*. 1071–1079.

[11] Li Li, Yisheng Lv, and Fei-Yue Wang. 2016. Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica* 3, 3 (2016), 247–254.

[12] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Joseph Gonzalez, Ken Goldberg, and Ion Stoica. 2017. Ray RLLib: A Composable and Scalable Reinforcement Learning Library. *arXiv preprint arXiv:1712.09381* (2017).

[13] Matthias Mueller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. 2018. Driving Policy Transfer via Modularity and Abstraction. In *Conference on Robot Learning*. IEEE, 1–15.

[14] Gábor Orosz. 2016. Connected cruise control: modelling, delay effects, and nonlinear behaviour. *Vehicle System Dynamics* 54, 8 (2016), 1147–1176.

[15] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2017. Sim-to-real transfer of robotic control with dynamics randomization. *arXiv preprint arXiv:1710.06537* (2017).

[16] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust adversarial reinforcement learning. *arXiv preprint arXiv:1703.02702* (2017).

[17] Jackeline Rios-Torres and Andreas A. Malikopoulos. 2017. A Survey on the Coordination of Connected and Automated Vehicles at Intersections and Merging at Highway On-Ramps. *IEEE Transactions on Intelligent Transportation Systems* (2017), 1066–1077.

[18] Andrei A Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. 2016. Sim-to-real robot learning from pixels with progressive nets. *arXiv preprint arXiv:1610.04286* (2016).

[19] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International Conference on Machine Learning*. 1889–1897.

[20] Steven E Shladover. 2017. Connected and Automated Vehicle Systems: Introduction and Overview. *Journal of Intelligent Transportation Systems* just-accepted (2017), 00–00.

[21] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354.

[22] Adam Stager, Luke Bhan, Andreas Malikopoulos, and Liuhui Zhao. 2018. A Scaled Smart City for Experimental Validation of Connected and Automated Vehicles. 51, 9 (2018), 130 – 135. 15th IFAC Symposium on Control in Transportation Systems CTS 2018.

[23] D Swaroop and J Karl Hedrick. 1996. String stability of interconnected systems. *IEEE transactions on automatic control* 41, 3 (1996), 349–357.

[24] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 23–30.

[25] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical review E* 62, 2 (2000), 1805.

[26] Eugene Vinitsky, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M Bayen. 2018. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Conference on Robot Learning*. IEEE, 399–409.

[27] Zia Wadud, Don MacKenzie, and Paul Leiby. 2016. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. *Transportation Research Part A: Policy and Practice* 86 (2016), 1–18.

[28] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. 2017. Flow: Architecture and Benchmarking for Reinforcement Learning in Traffic Control. *arXiv preprint arXiv:1710.05465* (2017).

[29] Zhuo Xu, Chen Tang, and Masayoshi Tomizuka. 2018. Zero-shot Deep Reinforcement Learning Driving Policy Transfer for Autonomous Vehicles based on Robust Control. In *International Conference on Intelligent Transportation Systems*. IEEE, 2865–2871.