Real-World Autonomous Vehicle Control Trained Entirely within Data-Driven Simulation

Alexander Amini [1 2] Igor Gilitschenski [1] Jacob Phillips [1 2] Julia Moseyko [1 2] Sertac Karaman [3] Daniela Rus [1]

Abstract

Recent studies have shown that even vast collections of data from real drivers are insufficient to train autonomous vehicle controllers capable of generalizing to the variety of situations that can occur in the real world. End-to-end reinforcement learning within simulation presents many potential advantages to learn safety critical controller directly from an agent's raw perception. Unfortunately, existing simulators lack the photorealism needed to train such machine learning models for autonomous vehicles. In this work, we present a novel data-driven simulation and training engine capable of learning end-to-end autonomous vehicle controllers without any human supervision. We demonstrate the ability of these controllers to generalize to and navigate in the real world without access to any human control commands during training. Our results validate the learned control policy onboard a full-scale autonomous vehicle, including in previously un-encountered scenarios, such as new roads and novel, complex, near-crash situations.

1. Introduction

Deep learning has demonstrated remarkable performance in a diverse set of tasks, particularly in computer vision (Voulodimos et al., 2018) and natural language processing (Van Den Oord et al., 2016). When combined with reinforcement learning (RL), super-human performance can be achieved for planning and control tasks in structured environments, such as video or board games (Mnih et al., 2015; Silver et al., 2018). So far, this level of success has not been mirrored in the area of safety-critical control systems operating in unstructured real world environments such as autonomous cars. While end-to-end (i.e., perception-to-control) trained neural networks for autonomous vehicles have shown great promise (Pomerleau, 1989; Bojarski et al., 2016), they still lack methods to learn robust models at scale and require vast amounts of training data that is time consuming and expensive to collect. Even when trained with vast collections of data, supervised driving models only imitate human driving performance in similar situations, and do not generalize well to the complexities of real world driving, including varying road, weather, and traffic conditions. Capturing training data from all the necessary edge cases and non-ideal conditions, such as recovering from near collisions, is not only prohibitively expensive, it is also potentially dangerous (Kendall et al., 2018).

Training and evaluating autonomous vehicle controllers in simulation with synthetic data provides a potential solution to the need for more data and increased robustness to novel situations, and also avoiding the time, cost and safety is-

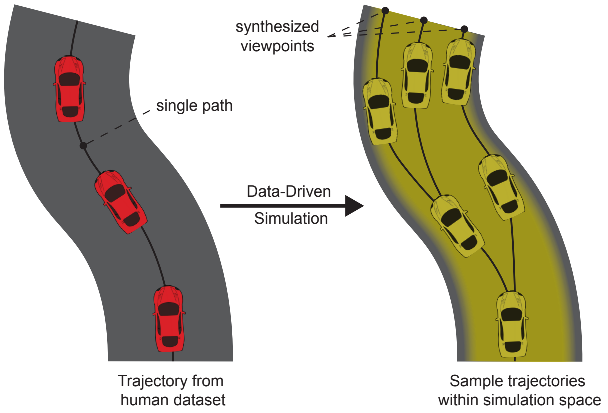
synthesized viewpoints
single path
Data-Driven Simulation
Trajectory from human dataset
Sample trajectories within simulation space

Figure 1. **TRAVIS: Training Robust Autonomous Vehicles in Simulation.** From a single human collected trajectory our data-driven simulator (TRAVIS) synthesizes a space of new possible trajectories. Preserving photo-realism of the real world allows the autonomous vehicle to move beyond imitation learning and instead explore the space, optimizing their own policies from scratch.

[1] Computer Science and Artificial Intelligence Lab (CSAIL), MIT [2] Department of Electrical Engineering and Computer Science (EECS), MIT [3] Laboratory of Information and Decision Systems (LIDS), MIT. Correspondence to: Alexander Amini <amini@mit.edu>.

Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning, Long Beach, California, USA, 2019. Copyright 2019 by the author(s).

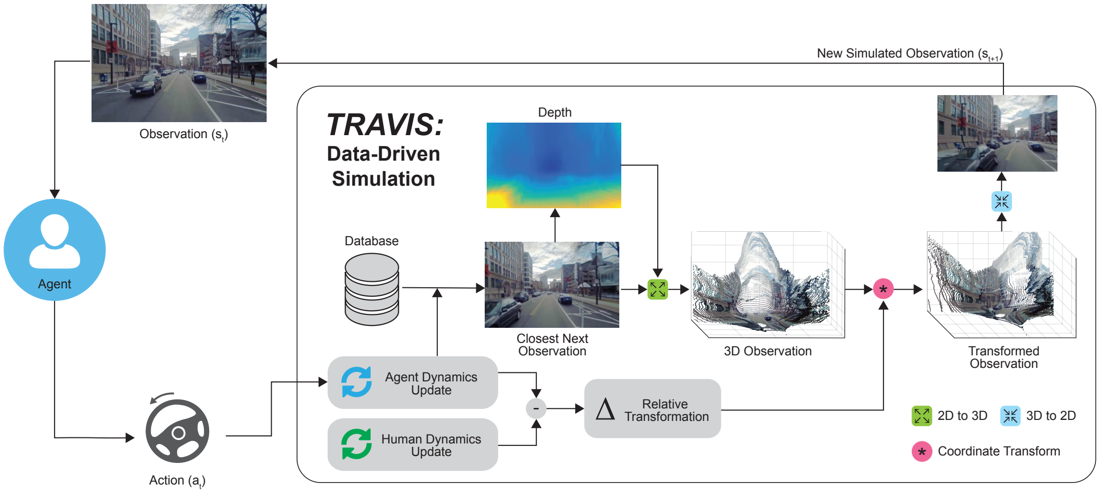

Figure 2. **Simulating novel viewpoints for learning.** Schematic of an autonomous agents interaction with the data-driven *TRAVIS* simulator. At every time step, the agent receives an observation of the environment and commands an action to execute. Their motion is simulated in *TRAVIS* and compared to the humans estimated motion in the real world. A new observation is then simulated by transforming a 3D representation of the scene into the virtual agents viewpoint.

sues of current methods. Unfortunately, existing simulators do not map well to the challenges of training end-to-end autonomous vehicle controllers.

We present an end-to-end simulation and training engine capable of taking a dataset of geo-tagged, human collected driving trajectories and synthesizing a continuum of new trajectories that are photorealistic and semantically faithful to the respective real world driving conditions (Fig. 1). Our engine, termed *TRAVIS: Training Robust Autonomous Vehicles in Simulation*, couples computer vision and deep learning techniques to synthesize this training data on-demand as a given trajectory is explored, and thus avoids the scalability issues of constructing an entire virtual world in advance. We consider the problem of learning lateral control for lane following over a wide variety of different road and environment types. Concretely, given only a visual observation of the environment (i.e. raw image pixels from a front facing video camera), we seek to learn a model that outputs the instantaneous curvature of the desired trajectory the vehicle should follow.

Our results show that RL agents trained entirely within *TRAVIS*, without any prior knowledge of human driving nor post-training fine-tuning, can be deployed directly onboard a full-scale autonomous vehicle capable of driving on real world roads. To the best of our knowledge, our work is the first published report of a full-scale autonomous vehicle trained entirely in simulation, using only reinforcement learning, that is capable of being deployed onto real roads and recovering from complex, near crash driving scenarios.

2. Related Work

Training agents in simulation for the purpose of robust deployment in the real world is a long-standing goal in many areas of robotics (Tobin et al., 2017; Andrychowicz et al., 2018; Sadeghi & Levine, 2016; Bewley et al., 2018). In autonomous driving, end-to-end trained controllers learn from raw perception data, as opposed to maps (Bansal et al., 2018) or other object representations (Chen et al., 2015; Henaff et al., 2019; Hong et al., 2018). Previous works have explored learning with expert information for lane following (Pomerleau, 1989; Bojarski et al., 2016), point-to-point navigation (Amini et al., 2019), and shared human-robot control (Amini et al., 2018), as well as in the context of reinforcement learning by allowing the vehicle to repeatedly drive off the road (Kendall et al., 2018). However, when trained on state-of-art synthetically generated images, these techniques are unable to be directly deployed in real world driving conditions. Thus, end-to-end trained controllers require photorealistic input data for robust training without any domain adaptation.

Performing style transformation, such as adding realistic textures to synthetic images, can be achieved with deep generative models and has been used to deploy learned policies from model-based simulation engines into the real world (Pan et al., 2017; Bewley et al., 2018). However, these approaches do not address the semantic complexities (such as driver and pedestrian behaviors) present in the real-world required to train robust autonomous controllers. Data driven engines like *The Gibson Environment* (Xia et al.,

2018) use a method of synthesizing photo-realistic environments, but such closed-world models are not scalable to the vast exploration space needed to train for real world autonomous driving. Simulation engines capable of training robust, end-to-end autonomous vehicle controllers must address the challenges of photo-realism, real world semantic complexities, and scalable exploration of control options, while avoiding the fragility of imitation learning and preventing unsafe conditions during data collection, evaluation, and deployment.

3. Data-Driven Simulation

In this section, we present our data-driven simulator, *TRAVIS*, and describe how autonomous vehicle controllers are trained via reinforcement learning on synthesized data to drive a stable trajectory consistent with the lane, allowing them to be deployed into real world.

TRAVIS is a data-driven simulator which is inserted into an agent's action-perception loop to synthesize photo-realistic local viewpoints as a virtual agent moves through the environment (Fig. 2). Here, we use the word "environment" in the reinforcement learning setting, where agents receive visual inputs and command a subsequent action to execute at that instant, while the simulator takes these commands to synthesize the next observation that the agent would observe. For each human-collected trajectory through a road environment, *TRAVIS* allows autonomous agents to drive (in virtual space) along an infinity of new local trajectories consistent with the road appearance and semantics, each with a different view of the scene.

The local viewpoint simulation occurs in three main stages. First, the action is used to update an internal state representation of the virtual agent and relative transformation from the human driver, inside the environment. Second, a new observation is retrieved from the database and projected from the sensor frame into the 3-dimensional (3D) world frame. The 3D observation then undergoes a coordinate transformation to account for the relative transformation between agent and human. Finally, the third stage of our system projects the transformed 3D observation back into the sensor frame of the vehicle and returns it to the agent as its next observation.

TRAVIS is scalable as it does not require storing and operating on 3D reconstructions of entire environments or cities. Instead, it considers only the observation collected nearest to the virtual agent's current state. Thus, simulating virtual agents over real road networks spanning thousands of kilometers can be achieved efficiently, with a few hundred gigabytes of data.

From only a single monocular image (Fig. 3A), a depth map is estimated using a deep convolutional neural network to handle objects on the road (e.g. cars, pedestrians, etc.) as well as off-road obstacles (e.g. signs, buildings, etc.). The monocular depth network is trained using self-supervision of stereo cameras, wherein the network accepts one camera and learns to predict the other camera by first learning a representation of the depth of pixels in the scene (Godard et al., 2017). *TRAVIS* is capable of simulating different local rotations (Fig. 3B) of the agent as well as both lateral (Fig. 3C) and longitudinal translations (Fig. 3D) along the road. We demonstrate simulated rotations up to $\pm 15^\circ$ as well as translations up to $\pm 1.5\mathrm{m}$. Since the average free lateral space of a vehicle within its lane is typically less than $1\mathrm{m}$, we demonstrate simulation beyond the bounds of lane-stable driving.

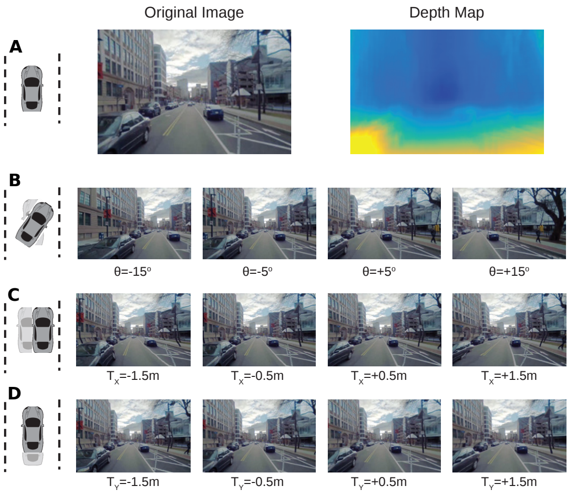

Figure 3. **Sample synthesized perception.** Example simulated images are shown for a sample environment observation and inferred depth map (A), including simulated rotations (B), lateral translations (C), and longitudinal translations (D).

3.1. End-to-end Learning

In this section, we present results on learning end-to-end (i.e., sensor-to-actuation) control of autonomous vehicles entirely within *TRAVIS*.

All controllers presented in this paper are learned end-to-end, directly from raw image pixels to actuation. We considered controllers that act based on their current perception without memory or recurrence built in, as suggested in (Bojarski et al., 2016; Chen et al., 2015). Features are extracted from the image using a series of convolutional layers that transform the image pixels into a lower dimensional feature space. These features are then fed through a set of dense fully connected layers to learn the final control commands to actuate the vehicle. Since all layers are fully differentiable, the model was optimized entirely end-to-end. As in previous work (Bojarski et al., 2016), we consider learning lateral control by predicting the curvature of motion that the vehicle should follow. Thus, our models remain vehicle independent, as they feed their desired output curvature into a vehicle-specific controller that computes a corresponding steering angle at deployment time.

Given a dataset of n observed state-action pairs $(s_t, a_t)_{i=1}^{n}$ from human driving, we aim to build an autonomous agent parameterized by $\boldsymbol{\theta}$ which estimates $\hat{a}_t = f(s_t; \boldsymbol{\theta})$. In the supervised learning setting, this agent outputs a *deterministic* action and is trained by minimizing the empirical error

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} (f(s_t; \boldsymbol{\theta}) - a_t)^2. \tag{1}$$

However, in the reinforcement learning setting, the agent has no explicit feedback of the human actuated command a_t. Instead, it receives a reward r_t for every consecutive action that does not result in a crash and can evaluate the return, $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, as the discounted, accumulated reward with a discount factor $\gamma \in (0, 1]$. In other words, the return that the agent receives at time t is a discounted distance traveled between t and the time when the vehicle crashes. As opposed to the supervised learning case, the agent optimizes a *stochastic* policy over the space of all possible actions: $\pi(a|s_t; \boldsymbol{\theta})$. Since the steering control of autonomous vehicles is a continuous variable, we parameterize the output probability distribution at time t as a Gaussian (μ_t, σ_t^2). Therefore, the policy gradient, $\nabla_{\boldsymbol{\theta}} \pi(a|s_t; \boldsymbol{\theta})$, of the agent can be computed analytically:

$$\nabla_{\boldsymbol{\theta}} \pi(a|s_t; \boldsymbol{\theta}) = \pi(a|s_t; \boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} \log \left(\pi(a|s_t; \boldsymbol{\theta}) \right) \tag{2}$$

Thus, the weights $\boldsymbol{\theta}$ are updating in the direction $\nabla_{\boldsymbol{\theta}} \log \left(\pi(a|s_t; \boldsymbol{\theta}) \right) \cdot R_t$ during optimization of the agent when training (Williams, 1992).

We present a reinforcement learning agent that learned to operate in various different simulated environments, where it only receives rewards based on how far it can drive without crashing. Compared to supervised learning, where agents learn to simply imitate the behavior of the human driver, reinforcement learning in simulation allows agents to learn suitable actions which maximize their total reward in that particular situation. Thus, the agent has no knowledge of how the human drove in that situation. Instead, it receives a reward the longer it stays safely on the road and is penalized when crashing off the edge of the lane boundary. Using only the feedback from crashes in simulation, the agent learns to optimize its own policy and thus to drive longer distances.

We define a learning episode in our simulator from the time the agent starts receiving sensory observations to the moment when it exits its lane boundaries. Assuming the original data was collected at approximately the center of the lane, this corresponds to declaring the end of an episode as when the lateral translation of the agent exceeds $\pm 1\mathrm{m}$. We train an convolutional neural network with a single Gaussian distribution output to model the continuous steering control of the vehicle in the simulator using policy gradient (PG) reinforcement learning (Williams, 1992; Sutton et al., 2000).

If the agent reaches the end of a road successfully, it is automatically transported to a new random location in the dataset. By doing so, we are not limited to simulating only on long roads, but can also train on multiple shorter roads as well. We define that an agent has sufficiently learned the environment once it successfully drives for 10km without crashing, at which point the simulator will restart without a crash penalty to continue with the next training iteration.

4. Results

4.1. Real-World Testbed

Learned controllers were deployed directly onboard a full-scale autonomous vehicle (2015 Toyota Prius V) which we retrofitted for full autonomous control (Naser et al., 2017) (Fig. 4). The primary perception sensor for control is a LI-AR0231-GMSL camera, which is recorded at 15Hz for training and simulation. The vehicle is also equipped with inertial measurement units (IMUs), wheel encoders, and a global positioning satellite (GPS) sensor for evaluation. The steering control commands from the models are sent to a low level tracking control running at 100Hz to actuate the physical steering wheel using the vehicle's built in power steering. To standardize all model trials on the test-track, a constant desired speed of the vehicle was set at 10 kph, while the model commanded the steering of the vehicle. All processing on the vehicle was done on an NVIDIA Drive PX2 computing platform with neural network inference using the on-board integrated Tegra GPUs running at a maximum of 30Hz.

Our test track contains a series of roads, turns, and intersec-

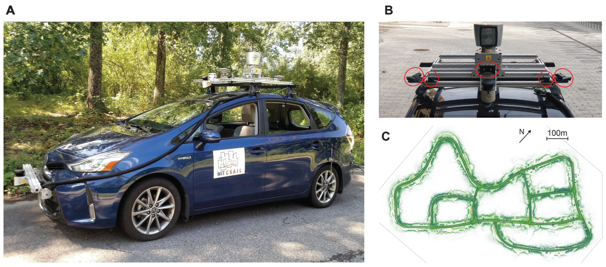

Figure 4. **Full-scale autonomous vehicle testbed.** Our algorithms and baselines are evaluated using a full-scale vehicle (A) on a real world test track. The vehicle is retrofitted with cameras (B, red circles). The test track (C) consists of rural roads without lane markers or clearly defined road boundaries, making perception an especially complex task.

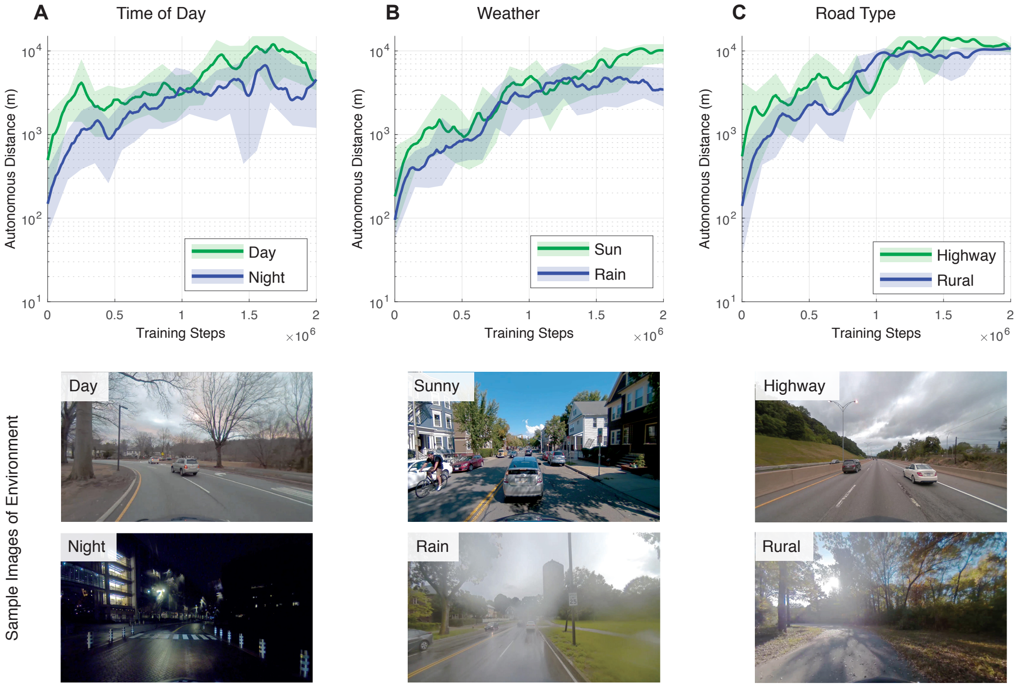

Figure 5. **Reinforcement learning in simulation.** Autonomous vehicles placed in the simulator with no prior knowledge of human driving or road semantics demonstrate the ability to learn and optimize their own driving policy under various different environment types. Scenarios range from different times of day (A), to weather condition (B), and road types (C).

tions spanning over 3 km. The track presents a difficult test environment, as it does not have any clearly defined road boundaries or lane markers. Cracks, where grass and other vegetation grow into the road, as well as strong shadows cast from surrounding trees, also make road detection significantly more challenging. Agents were evaluated on all roads in the test environment. It is crucial to note that the training set contained only the outermost loop of the track but in the reverse direction of the evaluation run. Additionally, all other side roads are also tested to evaluate generalization performance on entirely unseen roads.

4.2. Reinforcement Learning in TRAVIS

In this section, we present results on learning end-to-end (i.e., sensor-to-actuation) control of autonomous vehicles entirely within *TRAVIS*, under different weather conditions, times of day, and even road types. We evaluated training in different simulated environments synthesized from data collected by humans driving on real roads. Each of the environments collected for this experiment consisted of, on average, one hour of driving data from that scenario.

We started by learning end-to-end policies in different times of day (Fig. 5A) and, as expected, found that agents learned more quickly during the day than at night, where there was often limited visibility of lane markers and other road cues. Next, we considered changes in the weather conditions. Environments were considered “rainy” when there was enough water to coat the road sufficiently for reflections to appear, or when falling rain drops were visible in the images. Comparing dry with rainy weather learning, we found only minor differences between their optimization rates (Fig. 5B). This was especially surprising considering the visibility challenges for humans due to large reflections from puddles as well as raindrops covering the camera lens during driving. Finally, we evaluated different road types by comparing learning on highways and rural roads (Fig. 5C). Since highway driving has a tighter distribution of likely steering control commands (i.e., the car is traveling primarily in a nearly straight trajectory), the agent quickly learns to do well in this environment compared to the rural roads, which often have much sharper and more frequent turns. Additionally, many of the rural roads in our database lacked lane markers, thus making the beginning of learning harder since this is a key visual feature for autonomous navigation.

In our experiments, our learned agents iteratively explore and observe their surroundings (e.g. buildings, trees, cars, pedestrians, etc.) from novel viewpoints. On average, the learning agent is able to autonomously drive 10km without crashing in the environment within 1.5 million training iter-

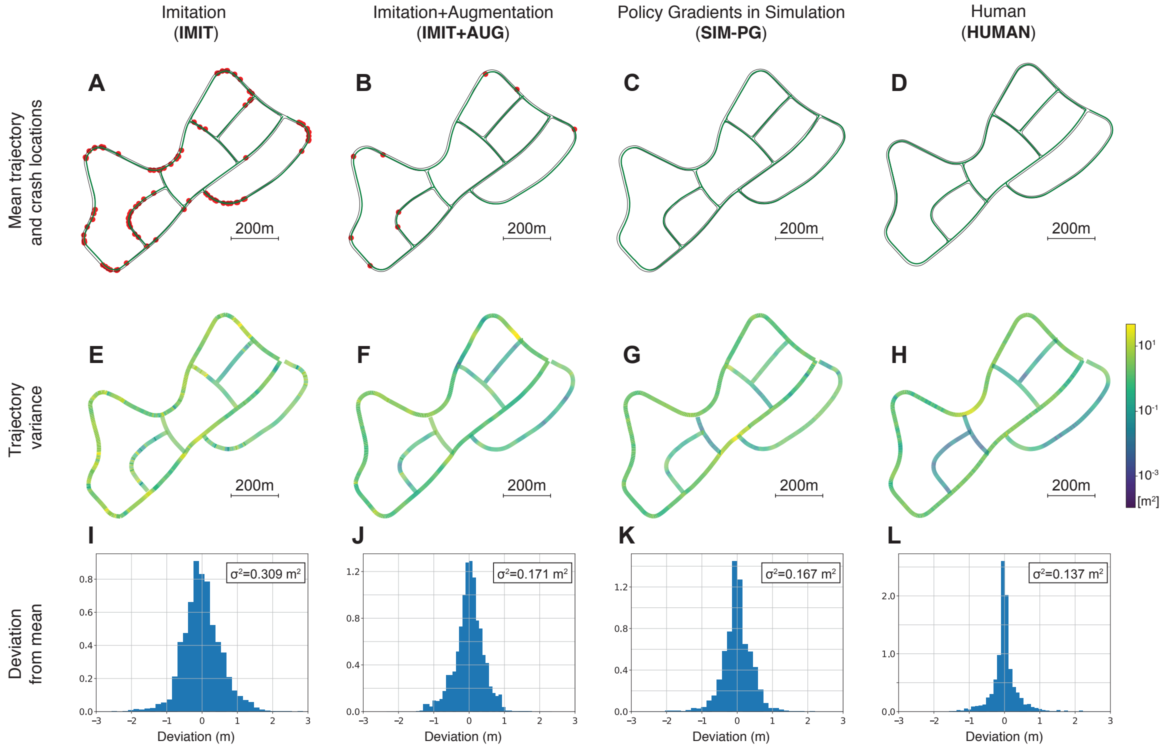

Figure 6. **Evaluation of end-to-end autonomous driving.** Comparison of different models on the test track including, standard imitation learning (**IMIT**), imitation learning with augmented off-center data (**IMIT-AUG**), policy gradient optimization within *TRAVIS* (**SIM-PG**), and baseline human driving. Each model is tested 3 times at fixed speeds on every road on the test track (A-D). Locations of crashes are indicated with red dots on the map. The mean trajectory of each model along with the variance between runs is also visualized (E-H). Finally, the distribution of deviations of the agent from its mean trajectory computed in (I-L).

ations. Thus, when randomly placed in new locations with similar features during training the agent is able to use its learned policy to navigate. Furthermore, while demonstration of learning in simulation is critical for development of autonomous vehicle controllers, we also evaluate the learned policies directly on-board our full-scale autonomous vehicle to test generalization to the real world.

4.3. Evaluation in the Real World

In this section, we evaluate the performance of learned agent policies directly on real roads. All policies were learned without human supervision, entirely within our data-driven simulator. Controllers trained within our data-driven simulator using policy gradient reinforcement learning (denoted by **SIM-PG**) are evaluated against two baseline supervised imitation learning techniques: (1) standard end-to-end supervised learning of road curvature (**IMIT**) and (2) supervised learning augmented with data from side cameras (**IMIT-AUG**). Augmenting supervised learning with views from side cameras is the standard approach to help teach the model how to recover from off-center positions on the roads, and has been used with great success (Bojarski et al., 2016; Giusti et al., 2016). We employ the techniques presented in (Bojarski et al., 2016) to compute the recovery correction signal that should be trained with given these augmented inputs. Finally, a human driver (**HUMAN**), is instructed to drive the designed route as close to the center of the lane as possible, and is used to fairly evaluate and compare all learned models.

Each of these models are trained separately three times and tested individually on every road on the test track. When reaching the end of a road, the autonomous vehicle is stopped and restarted at the beginning of the next road segment. The test driver intervenes and takes over control when the vehicle exits its lane or if it starts to go off-road. The mean trajectory of the three trials for each model are shown in Fig. 6A-D, with intervention locations visualized as red points. The road boundaries are plotted in black for scale of deviations. Both **IMIT** and **IMIT-AUG** experience multiple interventions over the course of the three trials, with **IMIT** requiring 70 interventions ($\sim$ 1 intervention every 120 meters). Since **IMIT** was not trained with any data from side views the car was unable to recover from off-center positions and thus drifts off the road even on minor turns. **IMIT-AUG** was more robust to slight off-center recoveries but still required intervention during large turns or when the vehicle encountered other edge-case orientations (e.g. large rotations on the road). **SIM-PG** exhibited the greatest robustness of the considered models and never required any interventions throughout the three different trials (totalling approximately 10km of autonomous driving).

The variance across trials is also visualized at each point along the trajectories of the models (Fig. 6E-H), where the color of the line maps onto the trial variance at that location. We note that for **IMIT** and **IMIT-AUG**, the variance tends to spike at locations which resulted in interventions, while the variance of **SIM-PG** was highest in ambiguous situations (such as approaching an intersection, or wider roads with multiple possible correct control outputs).

In addition to initiating testing of our autonomous controllers when the vehicle was started roughly in the center of the lane, we also tested initiating the vehicle from off-orientation positions, with significant lateral offsets as well as rotational offsets. Upon initiating the autonomous controller, we designated a successful recovery if the vehicle was able to successfully execute an evasive maneuver and drive back to the center of its lane within 5 seconds. Thus, this tests if the learned controllers are capable of recovering from these near-crash scenarios.

We observed that agents trained in *TRAVIS* (**SIM-PG**) were able to recover from these off-orientation positions on real and previously un-encountered roads, and also significantly outperformed models trained with imitation learning on real world data (**IMIT** and **IMIT-AUG**). Considering recovery from rotational offsets alone, **SIM-PG** successfully recovered over $3\times$ more frequently than **IMIT-AUG**. The performance of **IMIT-AUG** improved with translational offsets, but was still significantly outperformed by **SIM-PG** models trained in simulation by approximately 30%.

5. Conclusion

Simulation allows for scalable training of autonomous agents under a wide range of different large-scale environments as well as positions and orientations of the vehicle on the road. While there have been remarkable successes in learning real world controllers to imitate human drivers, these methods are often trained directly on data collected by human drivers. Collecting "gold-standard" human driving data to train imitation models can be extremely difficult and subjective from driver to driver. Additionally, due to safety concerns humans are limited in the types of scenarios that are feasible to collect (i.e., restricted to driving on the road and in safe conditions); thus, the resulting learned controllers rarely will be exposed to necessary edge case driving scenarios. Such edge cases are not only the most challenging to collect, but also the hardest for the controllers to robustly handle.

Our data-driven simulator supports training the controller anywhere within the feasible band of trajectories that can be synthesized from data collected by a human driver on a single trajectory. This enables training on infinitely more data from one or more individually collected trajectories. This allows for agents to move beyond imitation learning and to learn an entire driving policy from scratch by iteratively exploring the simulation space.

Our experiments empirically validate the ability to train models in our data-driven simulation engine using reinforcement learning, and to directly deploy these learned models on a full-scale autonomous vehicle that can then successfully drive autonomously on real roads. Furthermore, we demonstrate that controllers learned within our simulator exhibit greater robustness in recovery from near-crash scenarios. We believe our approach and system represents a major step towards the direct, real world deployment of end-to-end learning techniques for robust training of autonomous vehicle controllers.

References

Amini, A., Paull, L., Balch, T., Karaman, S., and Rus, D. Learning steering bounds for parallel autonomous systems. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2018.

Amini, A., Rosman, G., Karaman, S., and Rus, D. Variational end-to-end navigation and localization. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*, 2018.

Bansal, M., Krizhevsky, A., and Ogale, A. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

Bewley, A., Rigley, J., Liu, Y., Hawke, J., Shen, R., Lam, V.-D., and Kendall, A. Learning to Drive from Simulation without Real World Labels. *arXiv preprint arXiv:1812.03823*, 2018. URL http://arxiv.org/abs/1812.03823.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

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

Giusti, A., Guzzi, J., Ciresan, D., He, F.-L., Rodriguez, J. P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Di Caro, G., Scaramuzza, D., and Gambardella, L. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 2016.

Godard, C., Mac Aodha, O., and Brostow, G. J. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, pp. 7, 2017.

Henaff, M., Canziani, A., and LeCun, Y. Model-predictive policy learning with uncertainty regularization for driving in dense traffic. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HygQBn0cYm.

Hong, Z.-W., Yu-Ming, C., Su, S.-Y., Shann, T.-Y., Chang, Y.-H., Yang, H.-K., Ho, B. H.-L., Tu, C.-C., Chang, Y.-C., Hsiao, T.-C., et al. Virtual-to-real: Learning to control in visual semantic segmentation. *arXiv preprint arXiv:1802.00285*, 2018.

Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J.-M., Lam, V.-D., Bewley, A., and Shah, A. Learning to drive in a day. *arXiv preprint arXiv:1807.00412*, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Naser, F., Dorhout, D., Proulx, S., Pendleton, S. D., Andersen, H., Schwarting, W., Paull, L., Alonso-Mora, J., Ang, M. H., Karaman, S., et al. A parallel autonomy research platform. In *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pp. 933–940. IEEE, 2017.

Pan, X., You, Y., Wang, Z., and Lu, C. Virtual to real reinforcement learning for autonomous driving. 2017.

Pomerleau, D. A. ALVINN: An Autonomous Land Vehicle in a Neural Network. In *Advances in Neural Information Processing Systems 1*, pp. 305–313, 1989.

Sadeghi, F. and Levine, S. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. 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*, pp. 23–30. IEEE, 2017.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Xia, F., R. Zamir, A., He, Z.-Y., Sax, A., Malik, J., and Savarese, S. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.