Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **accepted version** of the book part:

Codevilla Moraes, Felipe; Santana, Eder; López Peña, Antonio M.; [et al.]. «Exploring the limitations of behavior cloning for autonomous driving». A: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019, p. 9328-9337. 10 pàg. Institute of Electrical and Electronics Engineers (IEEE). DOI 10.1109/ICCV.2019.00942

This version is available at https://ddd.uab.cat/record/274764

under the terms of the © IN COPYRIGHT license

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Exploring the Limitations of Behavior Cloning for Autonomous Driving

Felipe Codevilla [*]
Computer Vision Center (CVC)
Campus UAB, Barcelona, Spain
fcodevilla@cvc.uab.es

Eder Santana
Toyota Research Institute (TRI)
Los Altos, CA, USA.
edercsjr@gmail.com

Antonio M. López
Computer Vision Center (CVC)
Campus UAB, Barcelona, Spain
antonio@cvc.uab.es

Adrien Gaidon
Toyota Research Institute (TRI)
Los Altos, CA, USA.
adrien.gaidon@tri.global

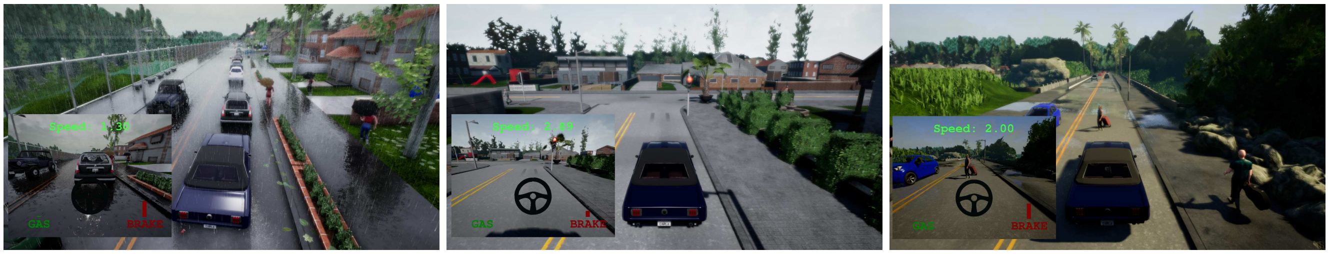

Figure 1. Driving scenarios from our new benchmark, based on the CARLA simulator, where the agent needs to react to dynamic changes in the environment, handle clutter (only part of the environment is causally relevant), and predict complex sensorimotor controls (lateral and longitudinal). We show that Behavior Cloning yields state-of-the-art policies in these complex scenarios and investigate its limitations.

Abstract

Driving requires reacting to a wide variety of complex environment conditions and agent behaviors. Explicitly modeling each possible scenario is unrealistic. In contrast, imitation learning can, in theory, leverage data from large fleets of human-driven cars. Behavior cloning in particular has been successfully used to learn simple visuomotor policies end-to-end, but scaling to the full spectrum of driving behaviors remains an unsolved problem. In this paper, we propose a new benchmark to experimentally investigate the scalability and limitations of behavior cloning. We show that behavior cloning leads to state-of-the-art results, executing complex lateral and longitudinal maneuvers, even in unseen environments, without being explicitly programmed to do so. However, we confirm some limitations of the behavior cloning approach: some well-known limitations (e.g., dataset bias and overfitting), new generalization issues (e.g., dynamic objects and the lack of a causal modeling), and training instabilities, all requiring further research before behavior cloning can graduate to real-world driving. The code, dataset, benchmark, and agent studied in this paper can be found at http://github.com/felipecode/coiltraine/blob/master/docs/exploring_limitations.md

1. Introduction

End-to-end behavior cloning for autonomous driving has recently attracted renewed interest [10, 8, 12, 43, 30] as a simple alternative to traditional modular approaches used in industry [13, 24]. In this paradigm, perception and control are learned simultaneously using a deep neural network. Explicit sub-tasks are not defined, but may be implicitly learned from data. These sensorimotor controllers are typically obtained by imitation learning from human demonstrations [2, 33, 1, 39]. The deep neural network learns, without being explicitly programmed, to recognize patterns associating sensory input (e.g., a single RGB image) with a desired reaction in terms of vehicle control parameters producing a target maneuver. Behavior cloning can directly learn from large fleets of human-driven vehicles without requiring a fixed ontology and extra manually labeled data. Finally, end-to-end imitation systems can be learned off-line in a safe way, in contrast to reinforcement learning approaches that typically require millions of trial and error runs in the target environment [25] or a faithful simulation.

End-to-end imitation systems can suffer a domain shift between the off-line training experience and the on-line behavior [35]. This problem, however, can be partially addressed in practice by data augmentation [8, 12]. Nonetheless, in spite of the early and recent successes of behavior cloning for end-to-end driving [32, 23, 10, 8, 12], it has not

[*] Work done during an internship at TRI.

yet proved to scale to the full spectrum of driving behaviors, such as reacting to multiple dynamic objects.

In this paper, we propose a new benchmark, called *NoCrash*, and perform a large scale analysis of end-to-end behavioral cloning systems in complex driving conditions not studied in this context before. We use a high fidelity simulated environment based on the open source CARLA simulator [14] to enable reproducible large scale off-line training and on-line evaluation in over 80 hours of driving under several different conditions. We describe a strong Conditional Imitation Learning baseline, derived from [12], that significantly improves upon state-of-the-art modular [26], affordance based [37], and reinforcement learning [27] approaches, both in terms of generalization performance in training environments and unseen ones.

Despite its positive performance, we identify limitations that prevent behavior cloning from successfully graduating to real-world applications. First, although generalization performance should scale with training data, generalizing to complex conditions is still an open problem with a lot of room for improvement. In particular, we show that no approach reliably handles dense traffic scenes with many dynamic agents. Second, we report generalization issues due to dataset biases and the lack of a causal model. We indeed observe diminishing returns after a certain amount of demonstrations, and even characterize a degradation of performance on unseen environments. Third, we observe a significant variability in generalization performance when varying the initialization or the training sample order, similar to on-policy RL issues [19]. We conduct experiments estimating the impact of ImageNet pre-training and show that it is not able to fully reduce the variance. This suggests the order of training samples matters for off-policy Imitation Learning, similar to the on-policy case [46].

Our paper is organized as follows. Section 2 describes related work, Section 3 our strong behavior cloning baseline, Section 4 our evalution protocol, including our new NoCrash benchmark, Section 5 our experimental results, and Section 6 our conclusion.

2. Related Work

Behavior cloning for driving dates back to the work of Pomerleau [32] on lane following, later followed by other approaches [23], including going beyond driving [1, 40]. The distributional shift between the training and testing distributions is the main known limitation of this approach, which might require *on-policy* data collection [34, 35], obtained by the learning agent. Nonetheless, recent works have proposed effective *off-policy* solutions, for instance by expanding the space of image/action pairs either using noise [22, 12], extra sensors [8], or modularization [37, 26, 5]. We show, however, that there are other limitations important to consider in complex driving scenarios, in particular dataset bias and high variance, which both harm scaling generalization performance with training data.

Dataset bias is a core problem of real-world machine learning applications [42, 6] that can have dramatic effects in a safety-critical application like autonomous driving. Imitation learning approaches are particularly sensitive to this issue, as the learning objective might be dominated by the main modes in the training data. Going beyond the original CARLA benchmark [14], we use our new NoCrash benchmark to quantitatively assess the magnitude of this problem on generalization performance for more realistic and challenging driving behaviors.

High variance is a key problem in powerful deep neural networks, and we show that high performance behavior cloning models are particularly suffering from this. This problem is related to sensitivity to both initialization and sampling order [31], reproducibility issues in Reinforcement Learning [19, 29], and the need to move beyond the i.i.d. data assumption towards curriculum learning [7] for sensorimotor control [46, 4].

Driving benchmarks fall in two main categories: off-line datasets, e.g., [15, 36, 44, 18], or on-line environments. We focus here on on-line benchmarks, as visuomotor models performing well in dataset-based evaluations do not necessarily translate to good driving policies [11]. Driving is obviously a safety-critical robotic application. Consequently, for safety and to enable reproducibility, researchers focus on using photo-realistic simulation environments. In particular, the CARLA open-source driving simulator [14] is emerging as a standard platform for driving research, used in [12, 30, 37, 27, 26]. Note, however, that transferring policies from simulation to the real-world is an open problem [28] out of the scope of this paper, although recent works have shown encouraging results [30, 45].

3. A Strong Baseline for Behavior Cloning

In this section, we first describe the behavior cloning framework we use, its limitations, and a robustified baseline that tries to tackle these issues.

3.1. Conditional Imitation Learning

Behavior cloning [32, 38, 35, 25] is a form of supervised learning that can learn sensorimotor policies from off-line collected data. The only requirements are pairs of input sensory observations associated with expert actions. We use an expanded formulation for self-driving cars called Conditional Imitation Learning, CIL [12]. It uses a high-level navigational command $\mathbf{c}$ that disambiguates imitation around multiple types of intersections. Given an expert policy $\pi^*(x)$ with access to the environment state x, we can execute this policy to produce a dataset, $D = \{\langle \mathbf{o}_i, \mathbf{c}_i, \mathbf{a}_i \rangle\}_{i=1}^N$, where $\mathbf{o}_i$ are sensor data observations, $\mathbf{c}_i$ are high-level commands (e.g., take the next right, left, or stay in lane) and $\mathbf{a}_i = \pi^*(x_i)$ are the resulting vehicle actions (low-level controls). Observations $\mathbf{o}_i = \{i, v_m\}$ contain a single image i and the ego car speed v_m [12] added for the system to properly react to dynamic objects on the road. Without the speed context, the model cannot learn if and when it should accelerate or brake to reach a desired speed or stop.

We want to learn a policy π parametrized by $\boldsymbol{\theta}$ to produce similar actions to π^* based only on observations $\mathbf{o}$ and high-level commands $\mathbf{c}$. The best parameters $\boldsymbol{\theta}^*$ are obtained by minimizing an imitation cost ℓ:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_i \ell\big(\pi(\mathbf{o}_i, \mathbf{c}_i; \boldsymbol{\theta}), \mathbf{a}_i\big). \tag{1}$$

In order to evaluate the performance of the learned policy $\pi(\mathbf{o}_i, \mathbf{c}_i; \boldsymbol{\theta})$ on-line at test time, we assume access to a score function giving a numeric value expressing the performance of the policy π on a given benchmark (cf. section 4).

3.2. Limitations

In addition to the distributional shift problem [35], behavior cloning presents some key limitations.

Bias in Naturalistic Driving Datasets. The appeal of behavior cloning lies in its simplicity and theoretical scalability, as it can indeed learn by imitation from large off-line collected demonstrations (e.g., using driving logs from manually driven production vehicles). It is, however, susceptible to dataset biases like all learning methods. This is exacerbated in the case of imitation learning of driving policies, as most of real-world driving consists in either a few simple behaviors or a heavy tail of complex reactions to rare events. Consequently, this can result in performance degrading as more data is collected, because the diversity of the dataset does not grow fast enough compared to the main mode of demonstrations. This phenomenon was not clearly measured before. Using our new *NoCrash* benchmark (section 4), we confirm it may happen in practice.

Causal Confusion. Related to dataset bias, end-to-end behavior cloning can suffer from causal confusion [16]: spurious correlations cannot be distinguished from true causes in observed training demonstration patterns unless an explicit causal model or on-policy demonstrations are used. Our new *NoCrash* benchmark confirms the theoretical observation and toy experiments of [16] in realistic driving conditions. In particular, we identify a typical failure mode due to a subtle dataset bias: the *inertia problem*. When the ego vehicle is stopped (e.g., at a red traffic light), the probability it stays static is indeed overwhelming in the training data. This creates a spurious correlation between low speed and no acceleration, inducing excessive stopping and difficult restarting in the imitative policy. Although mediated perception approaches that explicitly model causal signals like traffic lights do not suffer from this theoretical limitation, they still under-perform end-to-end learning in unconstrained environments, because not all causes might be modeled (e.g., some potential obstacles) and errors at the perception layer (e.g., missed detections) are irrecoverable.

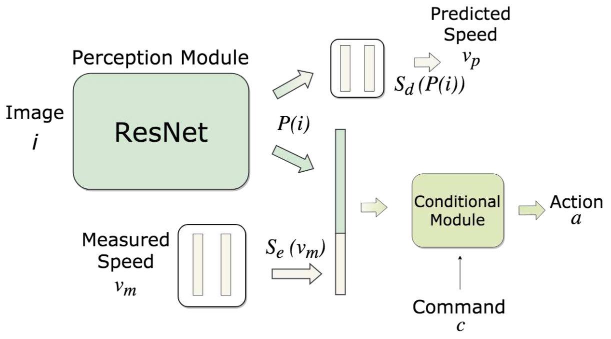

Figure 2. Our proposed network architecture, called CILRS, for end-to-end urban driving based on CIL [12]. A ResNet perception module processes an input image to a latent space followed by two prediction heads: one for controls and one for speed.

High variance. With a fixed off-policy training dataset, one would expect CIL to always learn the same policy in different runs of the training phase. However, the cost function is optimized via Stochastic Gradient Descent (SGD), which assumes the data is independent and identically distributed [9]. When training a reactive policy on snapshots of longer human demonstrations included in the training data, the i.i.d. assumption does not hold. Consequently, we might observe a high sensitivity to the initialization and the order in which the samples are seen during training. We confirm this in our experiments, finding an overall high variance due to both initialization and sampling order, following the decomposition in [31]:

$$\mathrm{Var}(\pi) = \mathbb{E}_D\big[Var_I(\pi|D)\big] + Var_D\big(\mathbb{E}_I[\pi|D]\big), \tag{2}$$

where I denotes the randomness in initialization. Because the policy π is evaluated on-line in simulated environments, we evaluate in practice the variance of the score on the test benchmark, and report results when freezing the initialization and/or varying the sampling order for different training datasets D (including of varying sizes).

3.3. Model

In order to explore the aforementioned limitations of behavior cloning, we propose a robustified CIL model designed to improve on [12] while remaining strictly off-policy. Our network architecture, called CILRS, is shown in Figure 2. We describe our enhancements below.

Deeper Residual Architecture. We use a ResNet34 architecture [17] for the perception backbone $\mathcal{P}(i)$. In the presence of large amounts of data, using deeper architectures can be an effective strategy to improve performance [17]. In particular, it can reduce *both* bias *and* variance, maintaining in particular a constant variance due to training set sampling with both network width and depth [31]. For end-to-end driving, the choice of architecture has been mostly limited to small networks so far [8, 12, 37] to avoid overfitting on limited datasets. In contrast, we notice that bigger models have better generalization performance on learning reactions to dynamic objects and traffic lights in complex urban environments.

Speed Prediction Regularization. To cope with the inertia problem without an explicit mapping of potential causes or on-policy interventions, we jointly train a sensorimotor controller with a network that predicts the ego vehicle's speed. Both neural networks share the same representation via our ResNet perception backbone. Intuitively, what happens is that this joint optimization enforces the perception module to have speed related features into the learned representation. This reduces the dependency on input speed as the only way to get dynamics of the scene, leveraging instead visual cues that are predictive of the car's velocity (e.g., free space, curves, traffic light states, etc).

Other changes. We use L1 as loss function ℓ instead of the mean squared error (MSE), as it is more correlated to driving performance [11]. As our *NoCrash* benchmark consists of complex realistic driving conditions in the presence of dynamic agents, we collect demonstrations from an expert game AI using privileged information to drive correctly (i.e. always respecting rules of the road and not crashing into any obstacle). Robustness to heavy noise in the demonstrations is beyond the scope of our work, as we aim to explore limitations of behavior cloning methods *in spite of* good demonstrations. Finally, we pre-trained our perception backbone on ImageNet to reduce initialization variance and benefit from generic transfer learning, a standard practice in deep learning seldom explored for behavior cloning.

4. Evaluation

In this section we discuss the simulated environment we use, CARLA, and review the original CARLA benchmark. Due to its limitations, we propose a new benchmark, called *NoCrash*, that tries to better evaluate driving controllers reaction to dynamic objects. This new benchmark, thanks to its complexity, allows further analysis on limitations of behavior cloning and other policy learning methods.

4.1. Simulated Environment

We use the CARLA simulator [14] version 0.8.4. The CARLA environment is divided in two different towns. Town 01 contains 2.9 km of drivable roads in a suburban environment. Town 02 is approximately 1.4 km of drivable roads, also in a suburban environment.

The CARLA environment may contain dynamic obstacles that interact with the ego car. Pedestrians, for instance, might cross the road on random occasions without any apparent previous notice. This action forces the ego car to promptly react. The CARLA environment also contains a diversity of car brands that cruise at different speeds. Overall it provides a diverse, photo-realistic, and dynamic environment with challenging driving conditions (cf. Figure 1).

The original CARLA benchmark [14] evaluates driving controllers on several goal directed tasks of increasing difficulty. Three of the tasks consist of navigation in an empty town and one of them in a town with a small number of dynamic objects. Each task is tested in four different conditions increasingly different from the training environment. The conditions are: same as training, new weather conditions that are derivatives from those seen during training, and a new town that has different buildings and different shadow patterns. Note that the biggest generalization test is the combination of new weather and new town.

The goal directed tasks are evaluated based on success rate. If the agent reaches the goal regardless of what happened during the episode, this episode is considered a success. The collisions and other infractions are considered and the average number of kilometers between infractions is measured. This evaluation induces the benchmark to be mainly focused on problems of a static nature. These problems consider the environmental conditions and the static objects of the world like buildings and trees. Thus, the original CARLA benchmark mostly evaluates skills such as lane keeping and performing 90 degrees turns.

4.2. NoCrash Benchmark

We propose a new larger scale CARLA driving benchmark, called *NoCrash*, designed to test the ability of ego vehicles to handle complex events caused by changing traffic conditions (e.g., traffic lights) and dynamic agents in the scene. For this benchmark, we propose different tasks and metrics than the original CARLA benchmark [14] to precisely measure specific reaction patterns that we know good drivers must master in urban conditions.

We propose three different tasks, each one corresponding to 25 goal directed episodes. In each episode, the agent starts at a random position and is directed by a high-level planner into reaching some goal position. The three tasks have the same set of start and end positions, as well as an increasing level of difficulty as follows:

1. Empty Town: no dynamic objects.
2. Regular Traffic: moderate number of cars and pedestrians.
3. Dense Traffic: large number of pedestrians and heavy traffic (dense urban scenario).

Similar to the CARLA Benchmark, *NoCrash* has six different weather conditions, where four were seen in training and two reserved for testing. It also has two different towns, one that is seen during training, and the other reserved for testing. For more details about the benchmark configuration, please refer to the supplementary material. As mentioned above, the measure of success of an episode should be more representative of the agent capabilities to react to dynamic objects. The original CARLA benchmark [14] has a goal conditioned success rate metric that is computed separately from a kilometers between infractions metric. The latter metric was proposed to be analogous to the one commonly used by real-world driving evaluations where the number of human interventions per kilometer is counted [20]. These interventions usually happen when the safety driver notices some inconsistent behavior that would lead the vehicle to a possibly dangerous state. On a potentially inconsistent behavior, the human intervention will put the vehicle back to a safe state. However, in the CARLA benchmark analysis, when an infraction is made, the episode continues after the infraction, leading to some inaccuracy in infraction counting. An example of inaccuracy includes whether a crash after leaving the road be counted as one or two infractions.

In *NoCrash*, instead of counting the number of infractions per kilometer, we end the episode as failing when any collision bigger than a fixed magnitude happens. With this limitation, we are setting a lower bound and have a guarantee of acceptable behaviors based on the measured percentage of success. Furthermore, this makes the evaluation even more similar to the km/interventions evaluation used in real world, since a new episode always sends the agent back to a safe starting state. In summary, we consider an episode to be successful if the agent reaches a certain goal under a time limit without colliding with any object. We also care about the ability of the agent to obey traffic rules. In particular, we measure and report the percentage of traffic light violations in Supplementary material. Note that an episode is not terminated when a traffic light violation occurs unless they are followed by a collision.

5. Experiments

In this section we detail our protocol for model training and briefly show that it is competitive with the state of the art. We also explore several corner cases to explore the limitations of the behavior cloning approach.

5.1. Training Details

First, we collected more than 400 hours of realistic simulated driving data from a single town of the CARLA environment using more than 200 GPU-days. We used an expert driving AI agent that leverages privileged information about the scene to drive naturally and well in complex conditions. After automatically filtering the data for simulation failures, duplicates, and edge cases using simple rules, we built a dataset of 100 hours of driving, called CARLA100. To enable running a wide range of experiments, we train all methods using a subset of 10 hours of expert demonstrations by default. We also report larger scale training experiments and scalability analyses in Section 5.3 and in supplementary material. One of the major differences of the training dataset, when compared to CIL, is that stopping for red traffic lights is considered on the demonstrator data. More details about the dataset are given in the supplementary material.

Training controllers on this dataset, we found that augmentation was not as crucial as reported by previous works [12, 26]. The only regularization we found important for performance was using a 50% dropout rate [41] after the last convolutional layer. Any larger dropout led us to underfitting models. All models were trained using Adam [21] with minibatches of 120 samples and an initial learning rate of 0.0002. At each iteration, a minibatch is sampled randomly from the entire dataset and presented to the network for training. If we detect that the training error has not decreased for over $1,000$ iterations we divide the learning rate by 10. We used a 2 hours validation dataset to discover when to stop the training process. We validate every 20k iterations and if the validation error increases for three iterations we stop the training process and use this checkpoint to test on the benchmarks, both CARLA and *NoCrash*. We build a validation dataset as described in [11].

5.2. Comparison with the state of the art

We compare our results using both the original CARLA benchmark from [14] and our proposed *NoCrash* benchmark. We compare two versions of our method: “CILRS” (our CIL extension with a ResNet architecture and speed prediction, as described in section 3), and a version without the speed prediction branch noted “CILR”. We compare our method with the original CIL from [12] and three state-of-the-art approaches: CAL [37], MT [26], and CIRL [27]. In contrast to end-to-end behavior cloning, these methods enforce some modularization that require extra information at training time, such as affordances (CAL), semantic segmentation (MT), or extra on-policy interaction with the environment (CIRL). Our approach only requires a fixed off-policy dataset of demonstrations.

We show results on the original CARLA benchmark [14] in Table 1 and results on our proposed *NoCrash* benchmark in Table 2. While most methods perform well in most conditions on the original CARLA benchmark, they all perform significantly worse on *NoCrash*, especially when trying to generalize to new conditions. This confirms the usefulness of *NoCrash* in terms of exploring the limitations of driving policy learning due to its more challenging nature.

In addition, our proposed CILRS model significantly improves over the state of the art, e.g., $+9\%$ and $+26\%$ on

	Training conditions						New town & weather					
Task	CIL[12]	CIRL[27]	CAL[37]	MT[26]	CILR	CILRS	CIL[12]	CIRL[27]	CAL[37]	MT[26]	CILR	CILRS
Straight	98	98	**100**	96	94	96	80	**98**	94	96	92	96
One Turn	89	**97**	**97**	87	92	92	48	80	72	82	**92**	**92**
Navigation	86	93	92	81	88	**95**	44	68	68	78	88	**92**
Nav. Dynamic	83	82	83	81	85	**92**	42	62	64	62	82	**90**

Table 1. Comparison with the state of the art on the original CARLA benchmark. The "CILRS" version corresponds to our CIL-based ResNet using the speed prediction branch, whereas "CILR" is without this speed prediction. These two models and CIL are the only ones that do not use any extra supervision or online interaction with the environment during training. The table reports the percentage of successfully completed episodes in each condition, selecting the best seed out of five runs.

	Training conditions					New Town & Weather				
Task	CIL[12]	CAL[37]	MT[26]	CILR	CILRS	CIL[12]	CAL[37]	MT[26]	CILR	CILRS
Empty	79 ± 1	81 ± 1	84 ± 1	92 ± 1	**97** ± 2	24 ± 1	25 ± 3	57 ± 0	66 ± 2	**90** ± 2
Regular	60 ± 1	73 ± 2	54 ± 2	72 ± 5	**83** ± 0	13 ± 2	14 ± 2	32 ± 2	54 ± 2	**56** ± 2
Dense	21 ± 2	**42** ± 3	13 ± 4	28 ± 1	**42** ± 2	2 ± 0	10 ± 0	14 ± 2	13 ± 4	**24** ± 8

Table 2. Results on our *NoCrash* benchmark. Mean and standard deviation on three runs, as CARLA 0.8.4 has significant non-determinism.

CARLA "Nav. Dynamic" in training and new conditions respectively, $+10\%$ and $+24\%$ on *NoCrash* Regular traffic in training and new conditions respectively. The significant improvements in generalization conditions, both w.r.t. CIL and mediated approaches, confirm that our improved end-to-end behavior cloning architecture can effectively learn complex general policies from demonstrations alone. Furthermore, our ablative analysis shows that speed prediction is helpful: CILR can indeed be up to -14% worse than CILRS on *NoCrash*.

5.3. Analysis of Limitations

Although clearly above the state of the art, our improved CILRS architecture nonetheless sees a strong degradation of performance similar to all other methods in the presence of challenging driving conditions. We investigate how this degradation relates to the limitations of behavior cloning mentioned in Section 3.2 by using the *NoCrash* benchmark, in particular to better evaluate the interaction of the agents with dynamic objects.

Generalization in the presence of dynamic objects. Limited generalization was previously reported for end-to-end driving approaches [14]. In our experiments, we observed additional, and more prominent, generalization issues when the control policies have to deal with dynamic objects. Table 2 indeed shows a large drop in performance as we change to tasks with more traffic, e.g., -55% and -66% from Empty to Dense traffic in *NoCrash* training / new conditions respectively. In contrast, results in Empty town only degrade by -7% when changing to a new environment and weather. Therefore, the learned policies have a much harder time dealing robustly with a large number of vehicles and pedestrians. Furthermore, this impacts all policy learning methods, including those using additional supervision or on-policy demonstrations, often even more than our proposed CILRS method.

Driving Dataset Biases. Figure 3 evaluates the effect of the amount of training demonstrations on the learned policy. Here we compare models trained with 2, 10, 50 and 100 hours of demonstrations. The plots show the mean success rate and standard deviation over four different training cycles with different random seeds. Our best results on most of the scenarios were obtained by using only 10 hours of training data, in particular on the "Dense Traffic" tasks and novel conditions such as New Weather and New Town.

These results quantify a limitation described in Section 3.2: the risk of overfitting to data that lacks diversity. This is here exacerbated by the limited spatial extent and visual variety of our environment, including in terms of dynamic objects. We indeed observed that some types of vehicles tend to elicit better reactions from the policy than others. The more common the vehicle model and color, the better the trained agent reacts to it. This raises ethical challenges in automated driving, requiring further research in fair machine learning for decision-making systems [6].

Causal confusion and the inertia problem. The main problem we observe caused by bias is the inertia problem stemming from causal confusion, as detailed in Section 3.2. Figure 4 shows the percentage of episodes that failed due to the agent staying still, without any intention to use the throttle, for at least 8 seconds before the timeout. Our results show the percentage of episodes failed due to that inertia problem increases with the amount of data used for training. We proposed to use a speed prediction branch as part of our CILRS model (cf. Figure 2) to mitigate this prob-

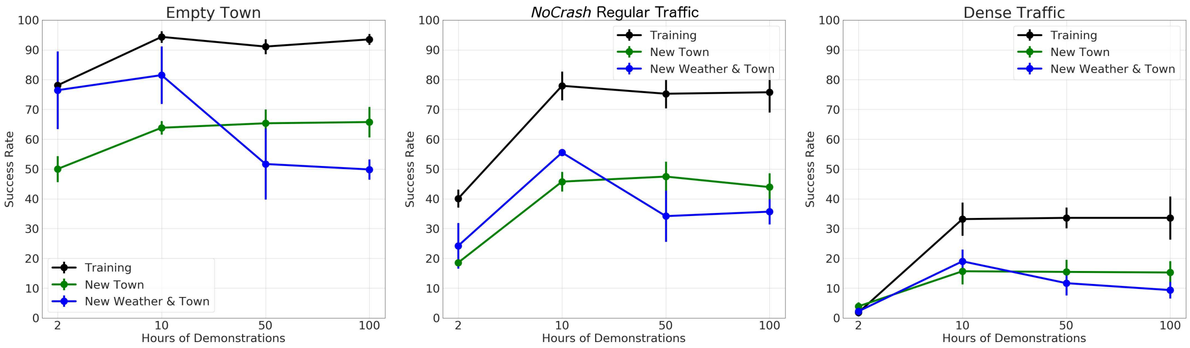

Figure 3. Due to biases in the data, the results may get either saturated or worse with increasing amounts of training data.

lem. Figure 5 shows the percentage of successes for the New Weather & Town conditions on different tasks with and without speed prediction. We observe that the speed prediction branch can substantially improve the success rate thanks to its regularization effect. It is, however, not a final solution to this problem, as we still observe instances of the inertia problem after using this approach.

High Variance. Repeatability of the training process is crucial for enhancing trust in end-to-end models. Unfortunately, we can still see drastic changes in the learned policy performance due to the variance caused by initialization and data sampling (cf. Section 3.2). Figure 6 compares the cause of episode termination for two models where the only difference is the random seed during training. The Model S1 has a much higher chance of ending episodes due to vehicle collisions. Qualitatively, it seemed to have learned a less general braking policy and was more prone to rear-end collisions with other vehicles. On the other hand, Model S2 is able to complete more episodes and is less likely to fail due to vehicle crashes. However, we can see that it times out more, showing a tendency to stop a lot, even in non threatening situations. This can be seen by analyzing the histograms of the throttle applied by both models during the benchmark, as shown in Figure 7. We can see a tendency for throttles of higher magnitude on Model S1.

As off-policy imitation learning uses a static dataset for training, this randomness comes from the order in which training data is sampled and the initialization of the random weights. This can possibly define which minima the models converge to. Table 3 quantifies the effect of initialization on the success rate of driving tasks by computing the variance expressed in Equation 2. The expected policy score was computed by averaging twelve different training runs. We also consider the variance with and without ImageNet initialization. We can see that the success rate can change by up to 42% for tasks with dynamic objects. ImageNet initialization tends to reduce the training variability, mainly due to smaller randomness on initialization but also due to a more stable learned policy.

	Task	Variance
CILRS	Empty	23%
	Regular	26%
	Dense	42%
CILRS (ImageNet)	Empty	4%
	Regular	12%
	Dense	38%

Table 3. Estimated variance of the success rate of CILRS on *NoCrash* computed by training 12 times the same model with different random seeds. The variance is reduced by fixing part of the initial weights with ImageNet pre-training.

6. Conclusion

Our new driving dataset (CARLA100), benchmark (*NoCrash*), and end-to-end sensorimotor architecture (CILRS) indicate that behavior cloning on large scale off-policy demonstration datasets can vastly improve over the state of the art in terms of generalization performance, including when comparing to mediated perception approaches with additional supervision. This is thanks to using a deeper residual architecture with an additional speed prediction target and good regularization.

Nonetheless, our extensive experimental analysis has shown that some big challenges remain open. First of all, the amount of dynamic objects in the scene directly hurts all policy learning methods, as multi-agent dynamics are not directly captured. Second, the self-supervised nature of behavior cloning enables it to scale to large datasets of demonstrations, but with diminishing returns (or worse) due to driving-specific dataset biases that require explicit treatment, in particular biases that create causal confusion (e.g., the inertia problem). Existing mitigation strategies currently need more informative intermediate representations, either learned [3] or using strong domain knowledge [5]. Third, the large variance resulting from initialization and sampling order indicates that multiple runs on the same off-

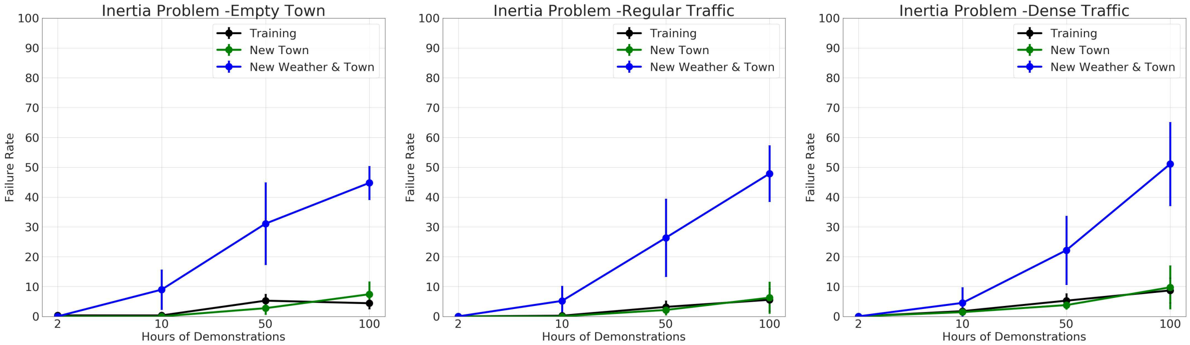

Figure 4. The percentage of episodes that failed due to the inertia problem. We can see that by increasing the amount of data, this bias may further degrade the generalization capabilities of the models.

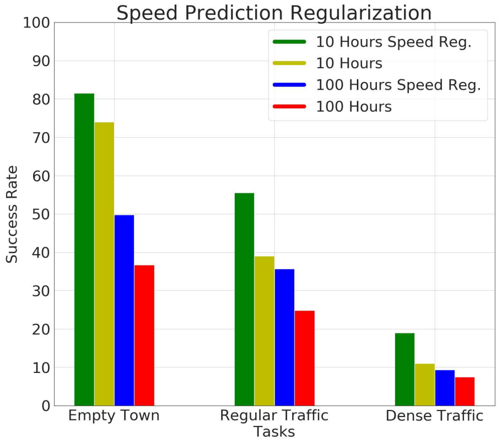

Figure 5. Comparison between the results with and without the speed prediction and different amounts of training demonstrations. We report the results only for the case were highest generalization is needed (New Weather and Town).

policy data is key to identify the best possible policies. This is part of the broader deep learning challenges regarding non-convexity and initialization, curriculum learning, and training stability.

Acknowledgements. Felipe Codevilla was supported in part by FI grant 2017FI-B1-00162. Antonio M. López acknowledges the financial support by ICREA under the ICREA Academia Program. Felipe and Antonio acknowledge the financial support by the Spanish TIN2017-88709-R (MINECO/AEI/FEDER, UE). As CVC/UAB researchers, they also acknowledge the Generalitat de Catalunya CERCA Program and its ACCIO agency. All authors thank the TRI support in AWS instances to run the additional experiments after Felipe's internship. Special thanks to CVC's PhD student Yi Xiao for her help on making the supplementary video.

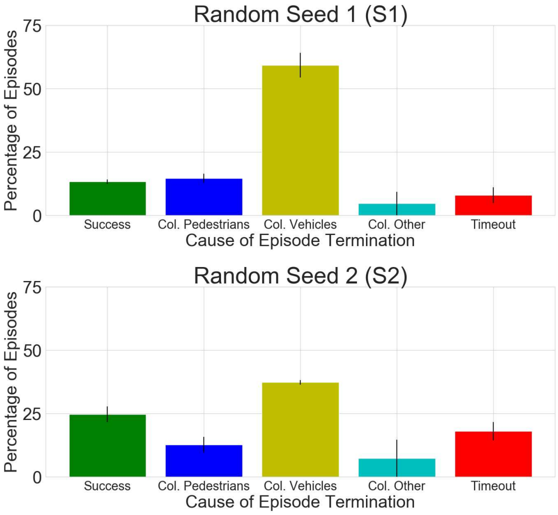

Figure 6. Cause of episode termination on *NoCrash* for two CILRS models (trained on 10 hours with ImageNet initialization) with identical parameters but different random seeds. The episodes were ran under "New Weather & Town" conditions of the "Dense Traffic" task.

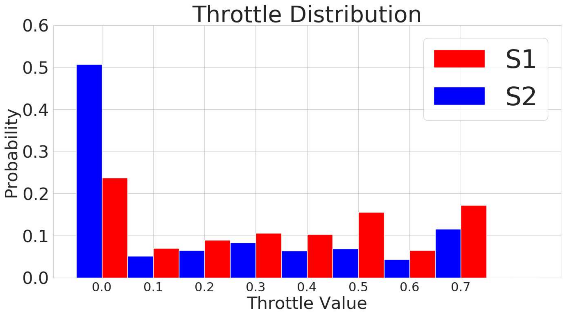

Figure 7. Probability distribution of having certain throttle values comparing models with two different random seeds but trained with the same hyper-parameters and data. We can see that S1 (red) is much more likely to have a higher throttle value.

References

[1] Pieter Abbeel, Adam Coates, Morgan Quigley, and Andrew Y. Ng. An application of reinforcement learn-

ing to aerobatic helicopter flight. In *NIPS*, 2006. 1, 2

[2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004. 1

[3] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Daniela Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *IROS*, 2018. 7

[4] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *NIPS*, 2017. 2

[5] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019. 2, 7

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. 2, 6

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 2

[8] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *arXiv:1604.07316*, 2016. 1, 2, 4

[9] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2008. 3

[10] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *ICCV*, 2015. 1

[11] Felipe Codevilla, Antonio M Lopez, Vladlen Koltun, and Alexey Dosovitskiy. On offline evaluation of vision-based driving models. In *ECCV*, 2018. 2, 4, 5

[12] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, 2018. 1, 2, 3, 4, 5, 6

[13] Ernst D Dickmanns. The development of machine vision for road vehicles in the last decade. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 1, pages 268–281. IEEE, 2002. 1

[14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 2, 4, 5, 6

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[16] P. Hamm, D. Jayaraman, and S. Levine. Causal confusion in imitation learning. In *”Neural Information Processing Systems Imitation Learning and its Challenges in Robotics Workshop (NeurIPS ILR)*, 2018. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[18] Simon Hecker, Dengxin Dai, and Luc Van Gool. End-to-end learning of driving models with surround-view cameras and route planners. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[19] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[20] Nidhi Kalra and Susan M. Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182 – 193, 2016. 5

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation (ICLR)*, 2015. 5

[22] Michael Laskey, Anca Dragan, Jonathan Lee, Ken Goldberg, and Roy Fox. Dart: Optimizing noise injection in imitation learning. In *Conference on Robot Learning (CoRL)*, 2017. 2

[23] Yann LeCun, Urs Muller, Jan Ben, Eric Cosatto, and Beat Flepp. Off-road obstacle avoidance through end-to-end learning. In *Neural Information Processing Systems (NIPS)*, 2005. 1, 2

[24] John Leonard, Jonathan How, Seth Teller, Mitch Berger, Stefan Campbell, Gaston Fiore, Luke Fletcher, Emilio Frazzoli, Albert Huang, Sertac Karaman, et al. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 25(10):727–774, 2008. 1

[25] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with large-scale data collection. In *International Symposium on Experimental Robotics (ISER)*, 2017. 1, 2

[26] Zhihao Li, Toshiyuki Motoyoshi, Kazuma Sasaki, Tetsuya Ogata, and Shigeki Sugano. Rethinking self-driving: Multi-task knowledge for better generaliza-

tion and accident explanation ability. *arXiv preprint arXiv:1809.11100*, 2018. 2, 5, 6

[27] Xiaodan Liang, Tairui Wang, Luona Yang, and Eric Xing. Cirl: Controllable imitative reinforcement learning for vision-based self-driving. In *ECCV*, 2018. 2, 5, 6

[28] A.M. Lopez, G. Villalonga, L. Sellart, G. Ros, D. Vzquez, J. Xu, J. Marin, and A. Mozafari. Training my car to see using virtual worlds. 2

[29] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018. 2

[30] Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladen Koltun. Driving policy transfer via modularity and abstraction. *arXiv preprint arXiv:1804.09364*, 2018. 1, 2

[31] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018. 2, 3, 4

[32] Dean Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Neural Information Processing Systems (NIPS)*, 1988. 1, 2

[33] Nathan D. Ratliff, James A. Bagnell, and Siddhartha S. Srinivasa. Imitation learning for locomotion and manipulation. In *International Conference on Humanoid Robots*, 2007. 1

[34] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668, 2010. 2

[35] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011. 1, 2, 3

[36] Eder Santana and George Hotz. Learning a driving simulator. *arXiv:1608.01230*, 2016. 2

[37] Axel Sauer, Nikolay Savinov, and Andreas Geiger. Conditional affordance learning for driving in urban environments. *arXiv preprint arXiv:1806.06498*, 2018. 2, 4, 5, 6

[38] Stefan Schaal, Auke Jan Ijspeert, and Aude Billard. Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society B*, 358(1431), 2003. 2

[39] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 2016. 1

[40] Sai Prashanth Soundararaj, Arvind K Sujeeth, and Ashutosh Saxena. Autonomous indoor helicopter flight using a single onboard camera. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5307–5314. IEEE, 2009. 2

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 5

[42] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE Computer Society, 2011. 2

[43] Q. Wang, L. Chen, and W. Tian. End-to-end driving simulation via angle branched network. *arXiv:1805.07545*, 2018. 1

[44] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[45] Luona Yang, Xiaodan Liang, Tairui Wang, and Eric Xing. Real-to-virtual domain unification for end-to-end autonomous driving. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[46] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *AAAI*, 2017. 2