Fighting Copycat Agents in Behavioral Cloning from Multiple Observations

Chuan Wen [*1] **Jierui Lin** [*2] **Trevor Darrell** [2] **Dinesh Jayaraman** [3] **Yang Gao** [2]

Abstract

Imitation learning trains policies to map from input observations to the actions that an expert would choose. In this setting, distribution shift frequently exacerbates the effect of misattributing expert actions to nuisance correlates among the observed variables. We observe that a common instance of this causal confusion occurs in partially observed settings when expert actions are strongly correlated over time: the imitator learns to cheat by predicting the expert's *previous* action, rather than the next action. To combat this "copycat problem", we propose an adversarial approach to learn a feature representation that removes excess information about the previous expert action nuisance correlate, while retaining the information necessary to predict the next action. In our experiments, our approach improves performance significantly across a variety of partially observed imitation learning tasks.

1. Introduction

Imitation learning [31, 35, 23, 25, 7, 14, 47] is a simple, yet powerful paradigm for learning complex behaviors from expert demonstrations. Given a training dataset of demonstrations, an agent's action policy π can be trained by mapping demonstration states s_t to expert actions a_t. Partially observed settings pose a problem for this approach: rather than the full state s_t, only observations o_t are available to the agent. However, this limitation may in principle be alleviated by a simple fix: instead of training a policy $\pi(a_t|o_t)$, one could train a policy $\pi(a_t|\tilde{o}_t = [o_t, o_{t-1}, o_{t-2}, \cdots])$, accessing past observations to fill in the missing state information of current observation.

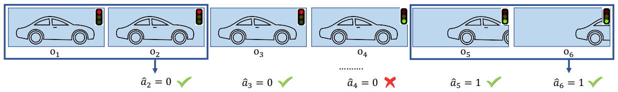

Figure 1: This figure demonstrates the "**copycat**" problem in an autonomous driving scenario. The top row is a sequence of observations, where the vehicle waits at the red light and start to drive when the light turns green. The policy takes a sliding window of observations as input. At the bottom row, we show that a "copycat" policy which simply replays its previous action will predict all but one actions correctly.

In practice, several prior works have reported that imitation from observation histories sometimes performs *worse* than imitation from a single frame alone [46, 23, 5]. To illustrate why this happens, consider the sequence of actions in an expert demonstration when it starts to drive in response to a red traffic light turning green (Figure 1). Assuming an action space with only two actions, brake ($a = 0$), and throttle ($a = 1$), the sequence of expert actions over time would look like $[a_0 = 0, a_1 = 0, \ldots, a_\tau = 0, a_{\tau+1} = 1, \ldots a_T = 1]$.

Which imitation policies would be effective at predicting the expert action on data from such demonstrations? Consider a "copycat" action policy that simply copies the previous expert action and prescribes repeating it as the next action. On these demonstrations, this policy would produce the correct action at all but one time instant, when the expert switches from braking to throttling. Our imitation learner $\pi(a_t|\tilde{o}_t)$ could easily expresss this copycat policy: it could recover the previous expert action a_{t-1} from the last two frames, o_t and o_{t-1}. The imitation training objective would encourage this, since this policy would produce low error on training and held-out demonstrations. However, when testing for actual driving performance, it would be useless — it would simply never switch to the throttle action.

We hypothesize that this copycat problem arises in imitation policies accessing past observations when two conditions are met: (i) expert actions over time are strongly correlated, and (ii) past expert actions are easily recovered from the observation history. As our first contribution, we empirically validate this hypothesis. We find that the temporal correlation among expert actions leads to even higher temporal correlation among learned policy actions. Further, the

[*] Equal contribution [1] IIIS, Tsinghua University, Beijing, China [2] EECS, UC Berkeley, California, USA [3] CIS, University of Pennsylvania, Pennsylvania, USA. Correspondence to: Yang Gao <yg@eecs.berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

higher this temporal correlation, the poorer the performance of the imitation learner (Section 3).

Then, we propose a novel imitation learning objective with an adversarial learning method to ensure the learned policy ignores the known nuisance correlate — the previous action a_{t-1}. Our implicit approach is scalable and robust, avoiding the need to learn disentangled representations [6], or learn a mixture of exponentially many graph conditioned policies [11]. Our method only needs an offline expert demonstration dataset, unlike methods like DAGGER [34] and CCIL [11] which resolve the causal confusion through online expert queries, or GAIL [18] which needs online environment interaction. Inspired by robotics applications, we demonstrate our approach in six simulated continuous robotic control settings.

2. Related Work

Imitation Learning. Imitation learning [27, 2], first proposed by Widrow and Smith in 1964 [48], enables the learning of complex behaviors from demonstrations. We focus on the widely used behavioral cloning paradigm [31, 35, 23, 25, 7, 14], which suffers from performance deteriorates when small errors in the learned policy compound over time. Existing solutions either assume access to a queryable expert as in DAGGER [34] and others [41, 22, 42] or refine the policy through environment interaction [18, 8].

de Haan et al. [11] explicitly connected distributional shift problems in imitation settings to nuisance correlations between input variables and expert actions, identifying the "causal confusion" problem. We isolate this causal confusion problem in its most frequently occurring form, the copycat problem motivated in Sec 1, encountered by ML practitioners within imitation learning [23, 5, 10, 46] and elsewhere, as "feedback loops" [37, 4]. We demonstrate a scalable solution to the copycat problem.

Causal Discovery. Models produced by standard machine learning approaches may rely on nuisance correlates, unlike *causal models* [29, 19, 30] that uncover relationships hold even under distributional shift. This connection between causality and distributional robustness has been studied in [17, 24, 9]. Existing causal discovery approaches, such as the widely used PC algorithm [39] and other techniques [6, 28, 16], operate over predefined, disentangled variables. However, in domains like vision or language, these underlying variables are unknown and hard to be inferred from raw, high-dimensional observations.

de Haan et al. [11] train a mixture of models on top of learned disentangled features in simple settings, exploiting environmental interactions afterwards to find the correct causal model. We take a different approach, sidestepping the difficulties of causal graph learning by injecting domain knowledge [26] to avoid nuisance correlates. Our approach is able to learn policies purely from demonstration data, requiring no environmental interaction afterwards.

Adversarial Learning. "Adversarial" learning approaches, set up to resemble two agents competing against each other, have recently had great success in many application doamins, such as image generation [15, 32, 20], and domain adaptation [45, 44, 12]. Our approach extends these adversarial losses to the imitation learning setting, setting up a nuisance correlate predictor as an adversary to the imitation policy.

3. The Copycat Problem

Conventional wisdom holds [27] that larger H, the size of the observation history window, would benefit the agent by providing more of the information contained in the state s_t. Yet, many prior works have reported [46, 23, 5, 10] that $H = 1$ is optimal. In other words, the *most poorly* observed setting, with $\tilde{o}_t = o_t$, yields the best results. Even more intriguingly, with $H > 1$, the likelihood $L(\theta^*)$ of held-out expert demonstrations improves, which means that there is no overfitting; only the environment reward $R(\theta^*)$ decreases. de Haan et al. [11] recently identified this problem as "causal confusion": BC policies misattribute expert actions to demonstration-specific nuisance correlates that no longer hold under the aforementioned distributional shift induced by policy execution.

We go one step further, pinning the nuisance correlate in a prominent class of causal confusion problems to the previous expert action a^e_{t-1}, which is often recoverable from observation histories, as in the car example in Sec 1. When expert actions are strongly correlated over time, the imitation learner has a suboptimal, but tantalizingly convenient shortcut [13]: merely learning to recover the previous action very nearly maximizes the training objective $L(\theta)$. We call this the copycat problem, and posit that it accounts for many reported cases of causal confusion.

Empirical Evidence for the Copycat Problem. We set the history size $H = 2$ and train a neural network policy $\pi_\theta(a|\tilde{o}_t = [o_t, o_{t-1}])$ on expert demonstrations. We call this policy behavior cloning with multiple observations (BC-MO). See Section 5 for more details.

To find a smoking gun for the copycat problem, we measure the predictability of the next action conditioned on the past actions, for a given policy. If the learner suffered from the copycat problem, this prediction would be easier on trajectories from the learned policy than on those from the expert policy. For each policy, we train a two-layer MLP to predict a_t given $a_{t-1}, \dots, a_{t-k}$ as input. Here we use $k = 9$. Table 1 shows the mean-squared error (MSE) on

Table 1: MSE for next action prediction, conditioned on previous actions. The lower the error for a policy, the higher its tendency to generate actions that can be predicted from previous actions alone.

	Ant	Hopper	Humanoid	Reacher	Walker2d	HalfCheetah
expert	0.0672	0.0091	0.746	0.000016	0.0257	0.0960
BC-MO	0.0067	0.0008	0.016	0.000004	0.0047	0.0269

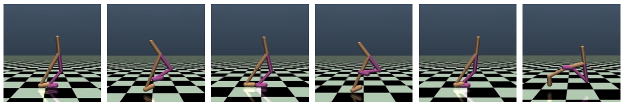

Figure 2: Frames of a Walker2D agent falling (left to right): the right knee repeats previous action at each time, failing to transition from an extended position to a bent position.

held-out trajectories, on all six MuJoCo environments. In each case, the MSE is lower for the cloned BC-MO policy, than for the expert, pointing to the copycat problem.

Figure 2 shows six frames of a copycat BC-MO agent falling in Walker2d, where the agent fails to switch its right knee from an extended to a bent position to maintain balance. Other cases are even worse: for example, we observed copycat Half-Cheetah agents that did not ever *start* to move from a resting position.

4. An Adversarial Solution to the Copycat Problem

We now propose an adversarial method to resolve the copycat problem. Our method builds on the standard behavioral cloning (BC) setup described in Sec 7.1. BC trains a policy $\pi_\theta(a_t|\tilde{o}_t)$ on expert demonstrations, to map from observation histories $\tilde{o}_t = [o_t, o_{t-1}, \cdots, o_{t-H+1}]$ to expert actions.

For notational simplicity, we now restrict ourselves to deterministic policies, so that $a_t = \pi_\theta(\tilde{o}_t)$. For a policy represented by a neural network, π_θ is easy to write as a composition of two learned functions, an encoder E and a decoder F: $a_t = \pi_\theta(\tilde{o}_t) = F(E(\tilde{o}_t))$. The output of the encoder network is a feature embedding $\boldsymbol{e_t} = E(\tilde{o}_t)$.

Adversarial Nuisance Variable Prediction. Fundamentally, the copycat problem arises from the fact that $\boldsymbol{e_t}$ contains information about the nuisance correlate a_{t-1}, and F learns to rely heavily on this information to predict a_t. This might suggest the following strategy: remove all information about a_{t-1} from the embedding $\boldsymbol{e}_t$, so that F cannot rely on a_{t-1} at all. In other words, we would train the encoder E to maximize the conditional entropy $H(a_{t-1}|\boldsymbol{e}_t)$, of the previous action, conditioned on the feature embedding. In practice, this means training an adversarial network D to predict a_{t-1} from $\boldsymbol{e}_t$.

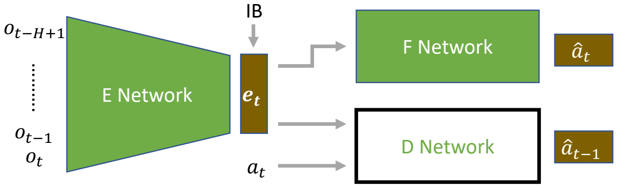

Figure 3: The network architecture. See Appendix for explainations.

However, note that removing all information about a_{t-1} may be counterproductive: after all, the copycat problem arises only when a_t and a_{t-1} are highly correlated. Removing all information about a_{t-1} would make it very difficult to predict a_t.

Target-Conditioned Adversary (TCA). To account for this, we design a slightly different adversarial strategy. Rather than removing all information about a_{t-1} from $\boldsymbol{e}_t$, we would like to remove only the information about a_{t-1} that is *not shared with* a_t. How might we set up an adversarial optimization process that removes this unshared information about a_{t-1}, while having no incentive to remove information that is actually useful for predicting the target variable a_t? The solution is simple: the adversary D still tries to predict a_{t-1} from $\boldsymbol{e}_t$, but with the target a_t as an additional conditioning input. The resulting optimization would train the encoder E to maximize the conditional entropy $H(a_{t-1}|\boldsymbol{e}_t, a_t)$.

Information Bottleneck (IB). Further, we add an information bottleneck [1] between the encoder E and the decoder F to express the prior that any residual excess information in $\boldsymbol{e}_t$ should be ignored. Specifically, we modify E to predict the parameters $\boldsymbol{\mu}_{\boldsymbol{e}_t}$ and $\boldsymbol{\sigma}_{\boldsymbol{e}_t}$ of an independent normal distribution, from which $\boldsymbol{e}_t \sim p_E(\boldsymbol{e}_t) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{e}_t}, \boldsymbol{\sigma}_{\boldsymbol{e_t}})$ is sampled. To apply an information bottleneck, we penalize the KL divergence between this distribution and the unit normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Finally, D operates directly on $\boldsymbol{\mu}_{\boldsymbol{e}_t}$, but F sees the "noisy" sample $\boldsymbol{e}_t$. Intuitively, the IB implements a penalty for every bit that E transmits to F, encouraging it to transmit only the most essential information and ignore the nuisance correlate.

Finally, putting the target-conditioned adversary and the information bottleneck together, we have the following minmax optimization problem:

$$\min_{E,F} \max_{D} V(E,F,D) = \mathbb{E}_{\boldsymbol{e}_t \sim p_E} \mathcal{L}(F(\boldsymbol{e_t}), a_t) + \\ \lambda KL(p_E(\boldsymbol{e}_t) || \mathcal{N}(\mathbf{0}, \mathbf{I})) - \alpha \mathcal{L}(D(\boldsymbol{\mu}_{e_t}, a_t), a_{t-1}) \qquad (1)$$

where $\mathcal{L}$ is an appropriate regression loss, such as the mean squared error.

Table 2: Cumulative rewards per episode in partially observed (PO) environments. The top half of the table shows results in our offline imitation setting. The lower half shows methods that additionally interact with the environment, including accessing reinforcement learning rewards and queryable experts. CCIL cannot run on Ant and Humanoid because of their high-dimensional observations.

	PO-Ant	PO-Hopper	PO-Humanoid	PO-Reacher	PO-Walker2d	PO-HalfCheetah
BC-SO	1300 ±148	275 ±40	587 ±58	−79 ±5	363 ±86	−38 ±36
BC-MO	1750 ±146	293 ±83	565 ± 80	−64 ±4	592 ±124	820 ±60
Dropout-BC [5]	830 ±330	223 ±49	577 ±65	−80 ± 7	283 ±194	406 ±165
Ours w/o Adversary	2030 ±88	473 ±129	638 ±101	−70 ±3	962 ±189	**1260** ±68
Ours w/o TCA	1629 ±287	322 ±74	607 ±58	−55 ±3	**1310** ±333	795 ±398
Ours w/o IB	1970 ±107	683 ±132	**696** ±48	−57 ±4	929 ±266	**1260** ±44
Ours	**2150** ±34	**1086** ±262	671 ±61	**−54** ±4	1296 ±288	1250 ±42
RL Expert [36]	2348 ±5	1780 ±0	4963 ±47	−10 ±0	2428 ±2	1336 ±4
CCIL [11]	-	145 ±55	-	−68 ±6	474 ±134	714 ±132
DAGGER [34] (100 queries)	2090 ±34	978 ±216	688* ±47	−52 ±4	701 ±133	1080 ± 86
DAGGER [34] (1k queries)	2240 ±14	1120 ±203	812* ±99	−15 ±2	2170 ±174	1270 ± 22

* for Humanoid, we used 100k and 500k queries for DAGGER, instead of 100 and 1000.

5. Experiments

In this section, we conduct experiments to evaluate our method against a variety of baselines. We also qualitatively study our method in order to better understand the newly introduced algorithm. We compare our method against others on all six MuJoCo [43] control environments from Open AI Gym, including Ant, HalfCheetah, Hopper, Humanoid, Reacher and Walker2D. These tasks vary broadly in their state and action spaces, environmental dynamics, and reward structure. See Appendix for more details.

5.1. Baselines

We compare our method to the following baselines: BC-SO (behavioral cloning with $H = 1$), BC-MO (behavioral cloning with $H > 1$), Dropout-BC [5], CCIL [11], DAGGER [34] and RL Expert [36].

5.2. Results and Analysis

For each method, in each environment, we train five policies with varying random initializations, and report the mean and standard deviation of their cumulative rewards. Table 2 shows these results. Our full method performs best, or tied best across all six environments, among purely offline imitation methods. As expected, behavior cloning from single observation (BC-SO) performs poorly, due to partial observability. BC-MO, with multiple observations, helps to varying extents in five out of six environments, but still performs much worse than the RL expert that generated the imitation trajectories. Dropout-BC [5] was originally proposed and evaluated in a setting where the nuisance correlate corresponded to a single dimension in the input, which doesn't hold when the nuisance variable is a function of the high-dimensional past observations. It performs uniformly poorly across all tasks.

Among non-offline imitators, CCIL fails to achieve better rewards than our approach, even with additional environmental interaction and queryable experts. We believe it's caused by the difficulty of learning a disentangled representation to handle high-dimensional observations in our settings. DAGGER (100 queries) performs comparably with our purely offline method, and DAGGER (1000 queries) performs significantly better, approaching the performance of the RL expert in several settings. On Hopper and Humanoid, even the best imitators fall far short of the expert.

Comparing ablated variants of our approach in Table 2, the target-conditioned adversary (TCA) and the information bottleneck (IB) are both clearly important components. Our method without TCA performs the poorest out of our ablations in four out of six environments — recall that the unconditioned adversary could force the removal of too much important information from the learned embedding. Using only the information bottleneck without an adversary (Ours w/o Adversary) already yields significant improvements over BC-MO — we take this to mean that merely restricting the description length of the learned representation encourages dropping nuisance information. Ours w/o IB, which only uses TCA, does nearly as well as our full method on most tasks, suggesting that the target-conditioned adversary is the most important component of our approach.

6. Conclusion and Future Work

In this paper, we identify the copycat problem that commonly afflicts imitation policies learning from histories of observations. We systematically study this phenomenon with a set of diagnostic experiments, which show the existence of this problem in multiple environments. Finally, we propose a new adversarial mechanism to tackle this problem. Our method significantly alleviates the copycat problem for offline behavioral cloning from multiple observations, and even outperforms some existing online behavioral cloning methods that have additional access to the environment.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57 (5):469–483, 2009.

[3] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hk4_qw5xe.

[4] Drew Bagnell. Feedback in machine learning, 2016. URL https://www.youtube.com/watch?v=XRSvz4UOpo4.

[5] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

[6] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

[7] 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. *CoRR*, abs/1604.07316, 2016.

[8] Kiante Brantley, Wen Sun, and Mikael Henaff. Disagreement-Regularized Imitation Learning. *International Conference in Learning Representations*, pages 1–19, 2020.

[9] Peter Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

[10] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9329–9338, 2019.

[11] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, pages 11693–11704, 2019.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org, 2015.

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[14] Alessandro Giusti, Jérôme Guzzi, Dan C Cireşan, Fang-Lin He, Juan P Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1 (2):661–667, 2015.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] Anirudh Goyal, Alex Lamb, Shagun Sodhani, Jordan Hoffmann, Sergey Levine, Yoshua Bengio, and Bernhard Scholkopf. Recurrent independent mechanisms, 2020. URL https://openreview.net/forum?id=BylaUTNtPS.

[17] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

[18] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4565–4573. Curran Associates, Inc., 2016.

[19] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[22] Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. Dart: Noise injection for robust imitation learning. In *Conference on Robot Learning*, pages 143–156, 2017.

[23] Yann LeCun, Urs Muller, Jan Ben, Eric Cosatto, and Beat Flepp. Off-road obstacle avoidance through end-to-end learning. *Advances in Neural Information Processing Systems*, pages 739–746, 2005. ISSN 10495258.

[24] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[25] Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32(3):263–279, 2013.

[26] Austin Nichols. Causal inference with observational data. *The Stata Journal*, 7(4):507–541, 2007.

[27] T Osa, J Pajarinen, G Neumann, JA Bagnell, P Abbeel, and J Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2): 1–179, 2018.

[28] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. *35th International Conference on Machine Learning, ICML 2018*, 9(1): 6432–6442, 2018.

[29] Judea Pearl. *Causality: Models, reasoning, and inference, second edition*. Cambridge university press, 2011. ISBN 9780511803161. doi: 10.1017/CBO9780511803161.

[30] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, November 2017.

[31] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313, 1989.

[32] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[33] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[34] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *Journal of Machine Learning Research*, 15:627–635, 2011. ISSN 15324435.

[35] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6): 233–242, 1999.

[36] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[37] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. IEEE, 2014.

[38] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised MAP inference for image super-resolution. *CoRR*, abs/1610.04490, 2016. URL http://arxiv.org/abs/1610.04490.

[39] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[40] 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.

[41] Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. *34th International Conference on Machine Learning, ICML 2017*, 7:5090–5108, 2017.

[42] Wen Sun, J. Andrew Bagnell, and Byron Boots. TRUNCATED HORIZON POLICY SEARCH: DEEP COMBINATION OF REINFORCEMENT AND IMITATION. In *International Conference on Learning Representations*, 2018.

[43] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[44] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[46] Dequan Wang, Coline Devin, Qi-Zhi Cai, Philipp Krähenbühl, and Trevor Darrell. Monocular plan view networks for autonomous driving. In *IROS*, 2019.

[47] Sean Welleck, Kianté Brantley, Hal Daumé III, and Kyunghyun Cho. Non-monotonic sequential text generation. *arXiv preprint arXiv:1902.02192*, 2019.

[48] Bernard Widrow and Fred W Smith. Pattern-recognizing control systems, 1964.