Generalization in Video Games with PPO!

Anders Nikolai Fure Nielsen - s192288, Hlynur Árni Sigurjónsson - s192302, Thordur Pall Fjalarsson - s192309

In this deep reinforcement learning project we approach the problem of training an agent to play video games with proximal policy optimization. While these algorithms have been shown to be advantageous over alternatives, they can exhibit problems with generalization to unseen levels. This project is a smaller scale experimentally based exploration of ideas from literature to improve the learning algorithm.

Key questions

- Data augmentation has been seen to improve generalizability for, can we see the same effect when trained for a limited amount of steps and levels?
- Are different types of data augmentations advantageous different settings, given our restrictions?
- How does different types of network architecture impact of the effect of the augmentations, given our restrictions?

Approach

- For all experiments have a set number of training levels and steps, lower than what is typically found in literature.
- Train agents with combinations of network architecture and data augmentations for a limited set of games.
- Analyze impact of changes.

Method

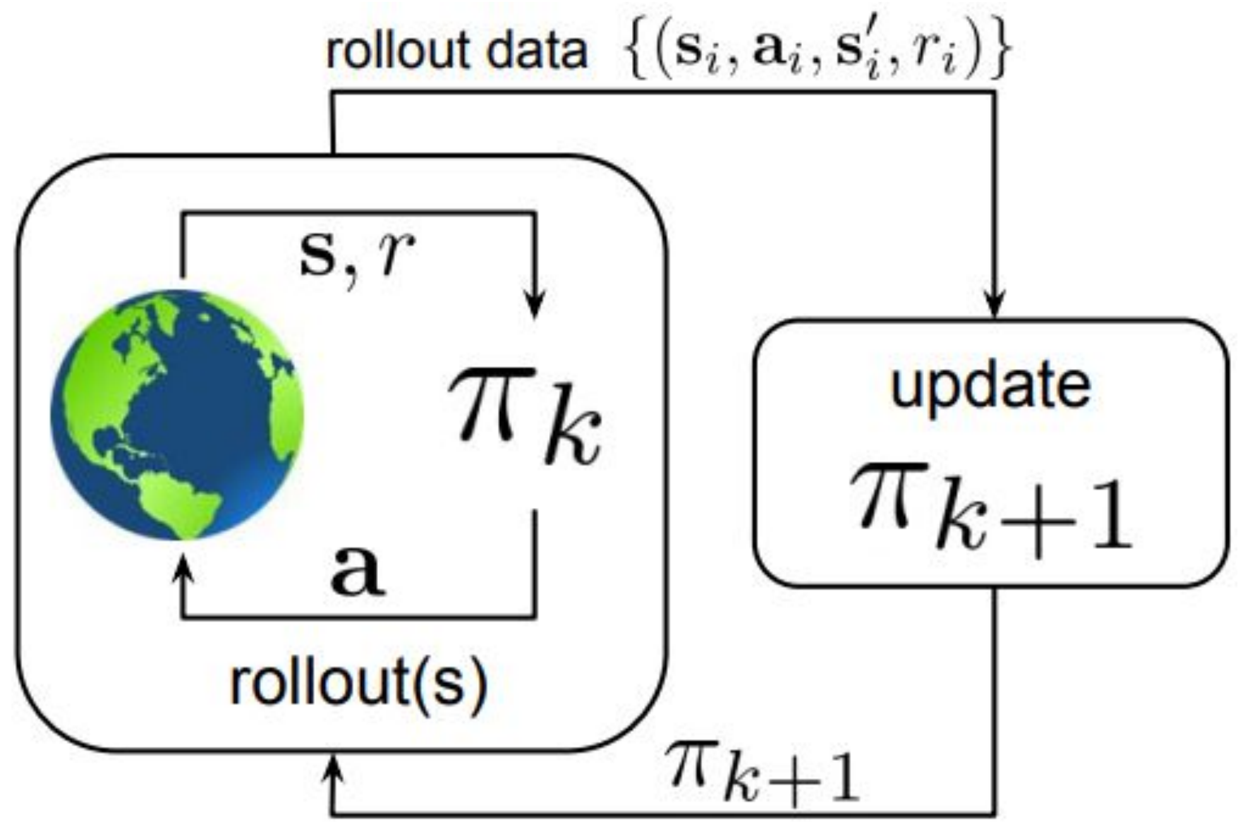
Algorithm: Proximal Policy Optimization (PPO)

$$\hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]$$

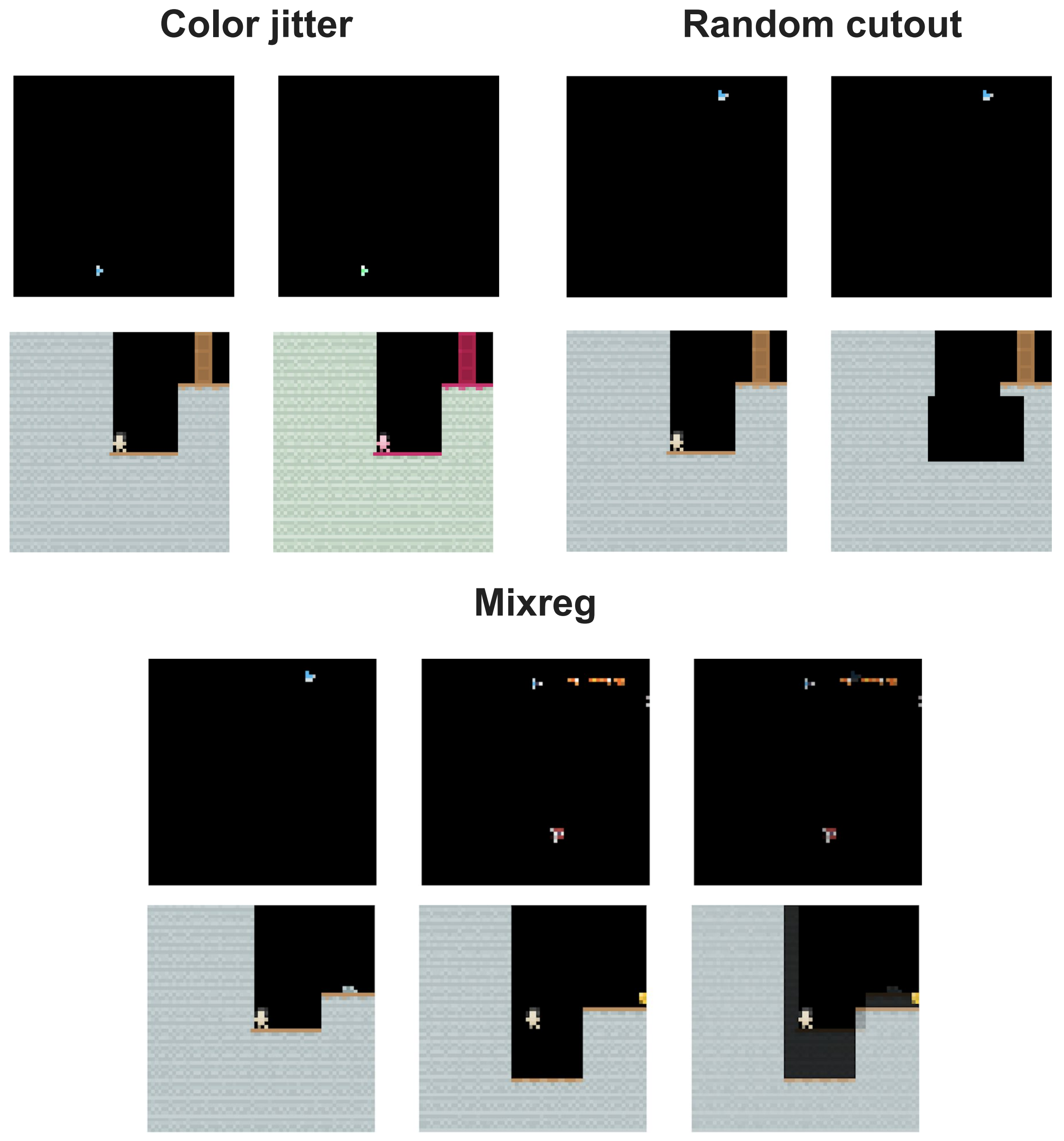
Encoder Networks: Impala Network & Nature network

Data Augmentation Schemes: Random Cutout, Random Color Jitter & Mixed Regularization

Online Reinforcement Learning

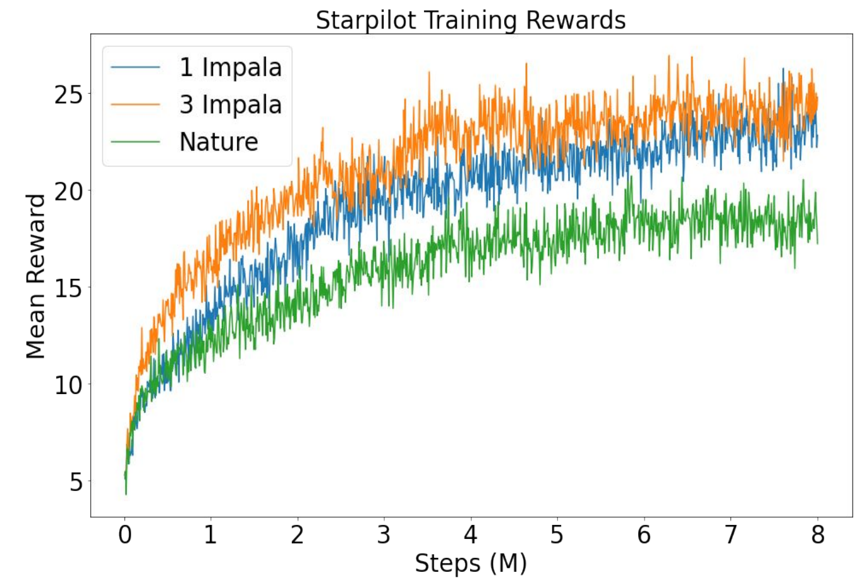


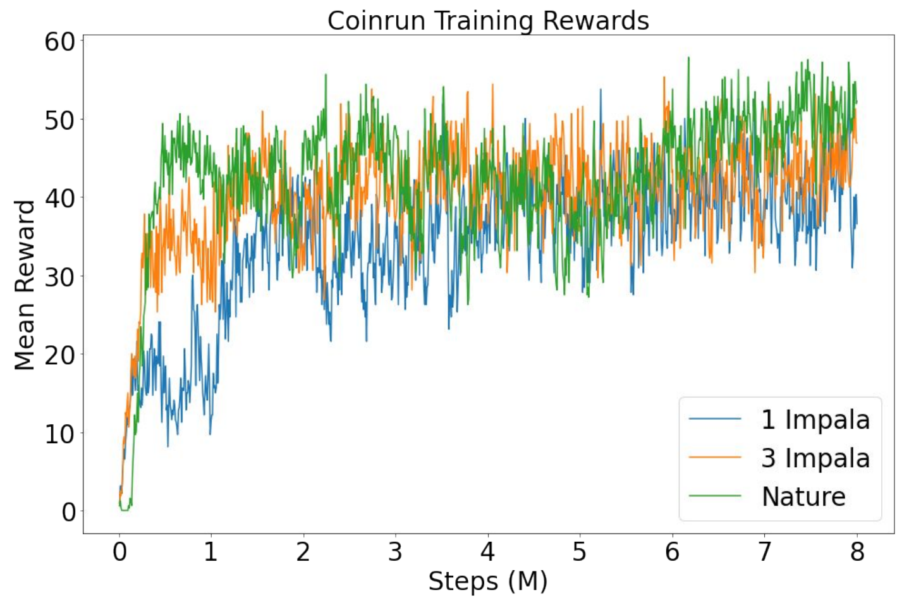
Data Augmentation Examples

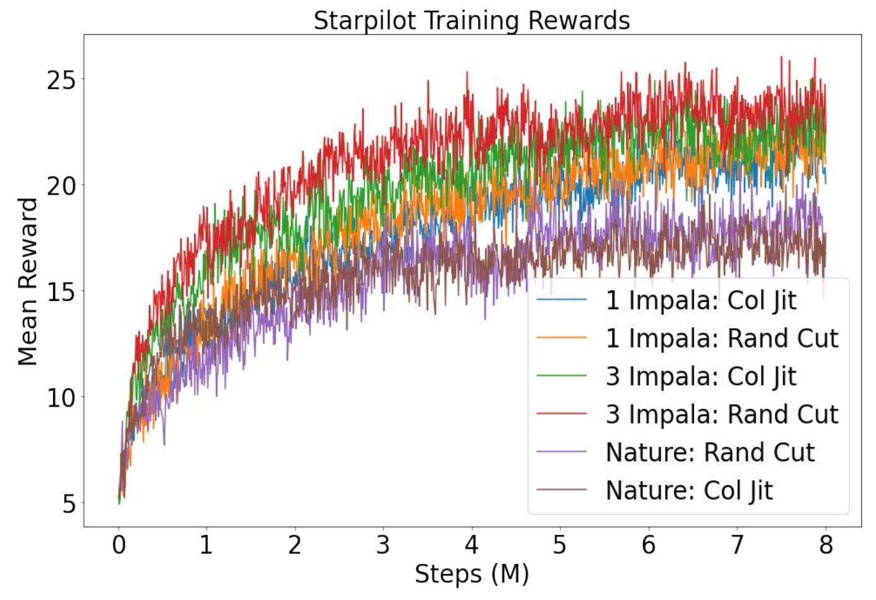


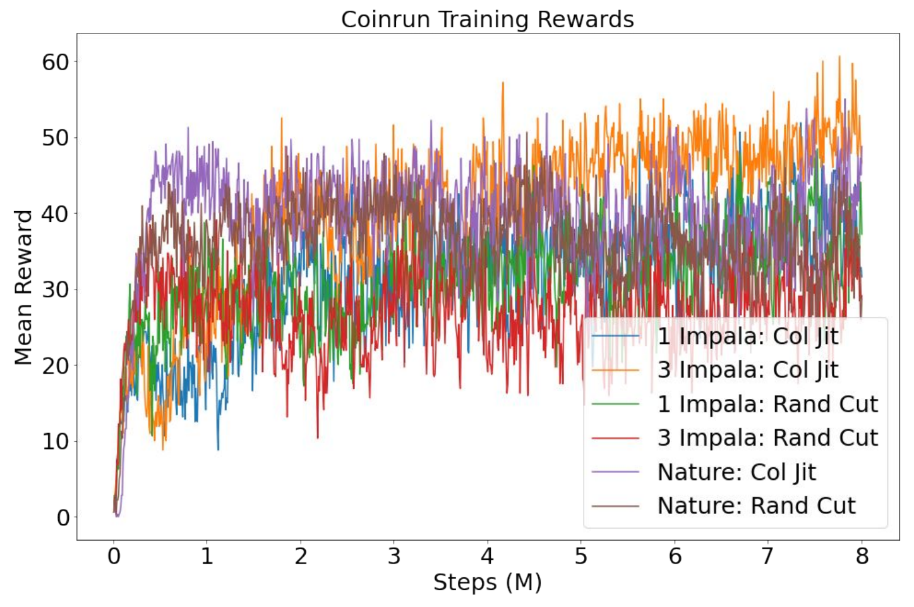
- Observations are randomly edited based on different augmentations.

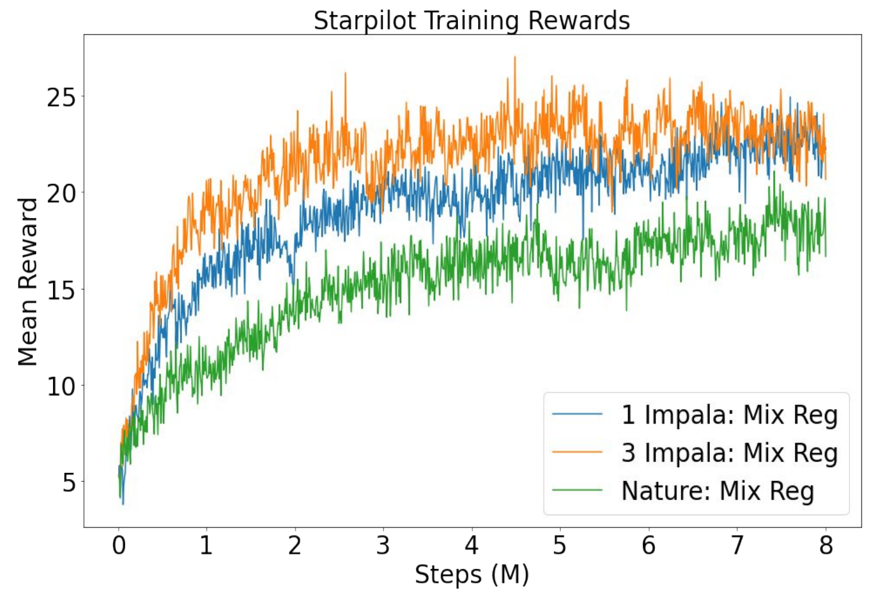
Training Performance

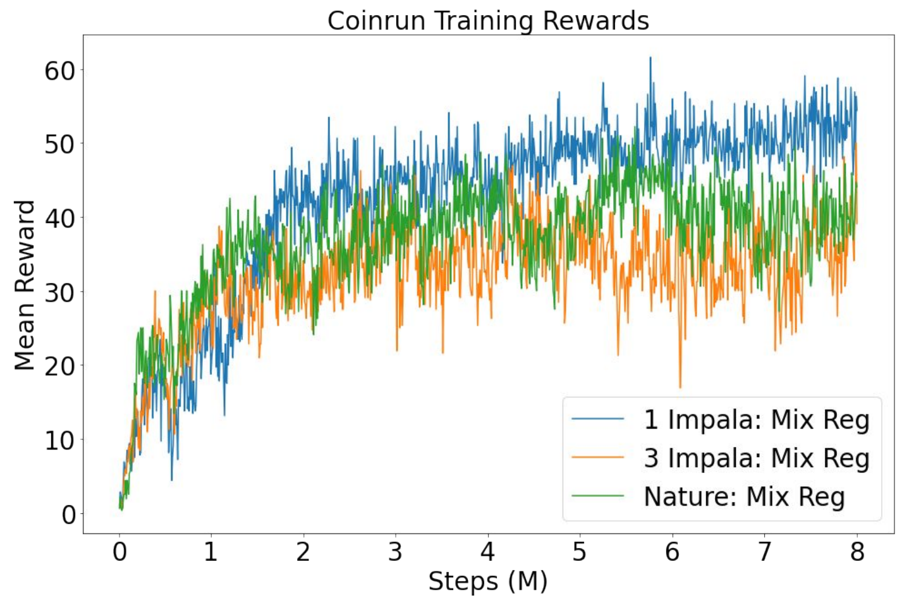












Test Performance

Table 1: Starpilot Mean Test Rewards

Experiment	Mean	Standard deviation
1 Impala: Col Jit	17.77	0.86
1 Impala: Mix Reg	18.93	0.59
1 Impala	18.32	0.34
1 Impala: Rand Cut	18.71	0.48
3 Impala: Col Jit	18.98	0.15
3 Impala: Mix Reg	19.04	0.45
3 Impala	19.60	0.64
3 Impala: Rand Cut	**19.73**	0.56
Nature: Rand Cut	18.19	0.59
Nature: Mix Reg	18.03	0.26
Nature: Col Jit	17.73	0.93
Nature	18.33	0.54

Table 2: Coinrun Mean Test Rewards

Experiment	Mean	Standard deviation
1 Impala	61.40	3.31
1 Impala: Rand Cut	54.60	0.78
1 Impala: Col Jit	55.44	1.61
1 Impala: Mix Reg	**63.93**	2.73
3 Impala	60.21	3.40
3 Impala: Rand Cut	56.20	2.06
3 Impala: Col Jit	61.96	2.92
3 Impala: Mix Reg	60.09	2.30
Nature	62.04	0.78
Nature: Rand Cut	52.48	2.09
Nature: Col Jit	57.45	2.97
Nature: Mix Reg	58.84	2.43

Key Takeaways

- We can observe learning, however differences in train performance does not necessarily reflect performance differences during testing.
- The effect of regularizing does not seem to have a consistent real effect on generalizability for our limited set of steps and levels.
- No approach is consistently better than the others for different games.
- Vastly different training run times.

References

[1] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[2] Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., & Srinivas, A. (2020). Reinforcement Learning with Augmented Data. *arXiv preprint arXiv:2004.14990*.

[3] Wang, K., Kang, B., Shao, J., & Feng, J. (2020). Improving Generalization in Reinforcement Learning with Mixture Regularization. Advances in Neural Information Processing Systems, 33.

[4] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.