A Minimalist Approach to Augmentation-based Self-Supervised Representation Learning for On-policy Reinforcement Learning

Nasik Muhammad Nafi
Kansas State University
Manhattan, KS, USA
nnafi@ksu.edu

William Hsu
Kansas State University
Manhattan, KS, USA
bshu@ksu.edu

ABSTRACT

Data augmentation has been proven as an effective measure to improve generalization performance in reinforcement learning (RL). Generic approaches directly use the augmented data to learn the value estimate or regularize the estimation, often ignoring explicit learning representation. On the other hand, algorithms that employ self-supervised representation learning mechanisms through data augmentation come at the cost of additional complexity. Such RL algorithms introduce new hyperparameters and design choices, often requiring additional components such as negative samples, projection heads, etc. We identify incorporating such mechanisms with on-policy RL algorithms poses additional challenges. In this work, we aim to develop a deep RL algorithm that ensures the benefits of both data augmentation and representation learning while incorporating minimal changes to the pre-existing on-policy RL algorithm. We find that we can match the performance of the state-of-the-art data augmentation-based self-supervised RL algorithms just by adding a simple non-contrastive loss with least required components. Our proposed approach **RAIR**: **R**einforcement learning with **A**ugmentation **I**nvariant **R**epresentation enables efficient representation learning and legitimate optimization sequence of objectives. We evaluate RAIR on all environments from the RL generalization benchmark Procgen. The experimental results indicate that RAIR outperforms PPO and achieves competitive or better performance with other data augmentation-based approaches.

KEYWORDS

Reinforcement learning, Representation learning, Non-contrastive, Data augmentation, Augmentation invariance, Generalization

ACM Reference Format:
Nasik Muhammad Nafi and William Hsu. 2025. A Minimalist Approach to Augmentation-based Self-Supervised Representation Learning for On-policy Reinforcement Learning. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Deep Reinforcement Learning (RL) agents trained directly from pixel-based observations often fail to generalize across different environmental conditions [1, 2, 8]. Training on diverse samples can improve generalization [2], however, in many real-world applications access to a large number of data can be costly and extremely difficult. Data augmentation offers an effective balance between the necessity and availability of the data. Existing RL methods integrate data augmentation in three broader ways: (i) naively applying augmented data for RL objective estimation [6], (ii) using additional losses for policy/value regularization that depends on augmented data [9, 13], or (iii) leveraging self-supervised contrastive or non-contrastive learning for improved representation learning [5, 7]. Learning from state-based representations is known to be more sample-efficient than from raw pixels [7]. However, we identify that such self-supervised loss-based approaches often exhibit instability due to their reliance on carefully designed components and extensive hyperparameter tuning, making them challenging to integrate with RL algorithms, particularly in online settings. Recent works reveal the potential of direct real-world RL training without pre-training on simulation [11, 12], where hyperparameter tuning is impractical, and sample efficiency and generalization are critical. Therefore, more stable, adaptable, and memory-efficient methods are needed for seamless real-world deployment of RL agents.

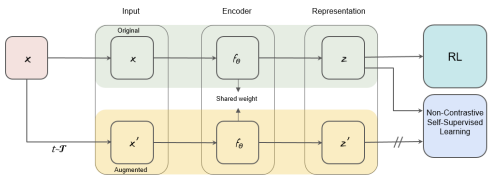

Figure 1: Overview of the proposed non-contrastive approach. We directly use the raw sample x for the RL objective and use the augmented one x' with stop-gradient only to generate pseudo labels for self-supervised representation learning.

In this paper, we propose Reinforcement learning with Augmentation Invariant Representation (RAIR), a minimalist approach that separates representation learning from the downstream RL task with minimal modifications to the underlying algorithm and architecture. We analyze design choices of self-supervised RL methods and streamline the framework to reduce inconsistency, hyperparameter complexity, and memory overhead. We validate our approach on all 16 Procgen environments [1] using widely used on policy actor-critic method Proximal Policy Optimization (PPO) [10].

2 METHODOLOGY

Figure 1 presents a high-level overview of our approach. We propose to use non-contrastive self-supervised learning for RL to ensure behaviorally similar states remain close in representation space,

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

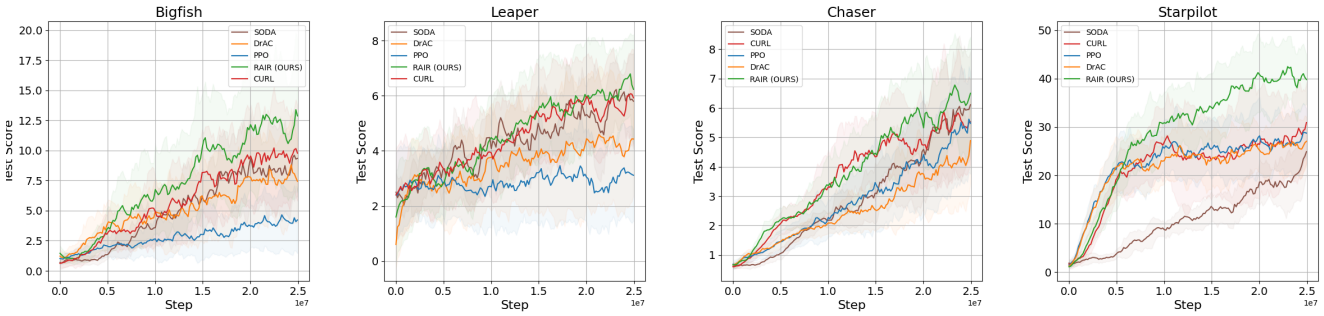

Figure 2: Test performance of RAIR (Ours), DrAC [9], CURL [7], SODA [5], and PPO [10] for "crop" (left two) and "color jitter" (right two) data augmentation in different Procgen environments. Mean and std are calculated over 5 trials.

Table 1: Comparison of components used in proposed RAIR

Method	Target Encoder	Projction Head	Predictor Head	Negative Sample
CURL	✓	×	×	✓
SODA	✓	✓	✓	×
RAIR	×	×	×	×

avoiding the risks of contrastive learning, which may incorrectly separate semantically similar states by considering them negative sample just due to coming from different trajectories.

Momentum encoders are commonly used in contrastive and non-contrastive learning to improve consistency and prevent representation collapse. However, we argue that in RL the concurrent optimization of the RL objective and representation loss inherently prevents collapse, making the use of momentum encoder redundant. Similarly, motivated by the computer vision research [4], projection and predictor heads are employed in non-contrastive RL to foster more expressive and diverse representations, aiming to prevent collapse to a single representation for all inputs [5]. We find that the RL objective, which is tangential to the representation learning objective, makes such additional heads (networks) redundant in the RL context. Removing these components decreases hyperparameters and reduces memory requirements associated with them. Table 1 presents a comparison of our proposed RAIR's architectural components as opposed to other existing methods.

To learn augmentation invariant representation, we learn encoder $f_\phi : \mathcal{S} \rightarrow \mathcal{Z}$ that captures similar representations for the original observation and its augmented counterpart. We train the encoder by minimizing the following objective:

$$J(\phi) = 1 - cos_s(z, \hat{z}) \tag{1}$$

where $z = f_\phi(s)$, $\hat{z} = f_\phi(t_r(s))$; and the state transformation mapping $t_r : \mathcal{S} \times \mathcal{H} \rightarrow \mathcal{S}$ denotes the augmentation function with $\mathcal{H}$ being the set of all possible parameters for $t_r(.)$. The cos_s refers to the cosine similarity metric. Further, we argue that in on-policy algorithms, updating the auxiliary representation learning loss after the RL objective (as in [5]) causes misalignment of the latent representation. To address this, we instead propose to optimize the auxiliary loss first, followed by the RL objective optimization.

Table 2: Average PPO-Normalized Return (%) of different approaches on test levels across all 16 environments. Evaluated with 5 trials for each environment, each with a different seed.

Augmentation	**PPO**	**DrAC**	**CURL**	**SODA**	**RAIR**
Crop	100.0	**136.0**	125.1	131.8	134.0
Color Jitter	100.0	119.3	138.3	129.8	**140.4**
Grayscale	100.0	95.0	**141.1**	136.8	130.9
Cutout	100.0	109.2	119.0	113.2	**127.4**

3 EXPERIMENTS AND DISCUSSIONS

We conduct our experiments on the full Procgen benchmark [1] consisting of sixteen procedurally generated environments. Each environment offers unlimited levels with diverse backgrounds and dynamics. We train the agent on 200 levels and test it on the full distribution to assess its generalization ability. We use the IMPALA-CNN architecture [3] for the encoder of RAIR and other baselines, where a fully connected layer maps the final CNN output to a 1024-dimensional latent representation. We compare RAIR against two relevant self-supervised approaches - a contrastive one, CURL [7] and a non-contrastive one, SODA [5]. Further, comparison with DrAC [9], an on-policy actor-critic approach, was conducted to validate robustness. Figure 2 and Table 2 shows that RAIR outperforms PPO and achieves competitive or better performance compared to other approaches. We also observe that RAIR performs consistently across different data augmentation techniques. Our code and long version of the paper are publicly available.[1]

4 CONCLUSION

RAIR presents a minimalist yet effective approach to augmentation-based RL, enabling competitive or superior generalization with minimal modifications to standard on-policy RL methods. RAIR directly aligns latent representations of augmented and original observations without requiring additional architectural components. Further, RAIR follows an appropriate order of updates that ensures consistency. Its simplicity makes it a compelling alternative to existing augmentation-based self-supervised RL techniques by offering efficiency, stability, and generalization for real-world applications.

[1] https://github.com/NasikNafi/RAIR

REFERENCES

[1] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*. PMLR, 2048–2056.

[2] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 2019. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1282–1289.

[3] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*. PMLR, 1407–1416.

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[5] Nicklas Hansen and Xiaolong Wang. 2021. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13611–13617.

[6] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. *Advances in neural information processing systems* 33 (2020), 19884–19895.

[7] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 5639–5650.

[8] Roberta Raileanu and Rob Fergus. 2021. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 8787–8798.

[9] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 5402–5415.

[10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[11] Laura Smith, Yunhao Cao, and Sergey Levine. 2024. Grow your limits: Continuous improvement with real-world rl for robotic locomotion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10829–10836.

[12] Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023. Demonstrating a walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. *Robotics: Science and Systems (RSS) Demo* 2, 3 (2023), 4.

[13] Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations*. https://openreview.net/forum?id=GY6-6sTvGaf