Reinforcement Learning with Augmentation Invariant Representation: A Non-contrastive Approach

Nasik Muhammad Nafi
Kansas State University
nnafi@ksu.edu

William Hsu
Kansas State University
bhsu@ksu.edu

Abstract

Data augmentation has been proven as an effective measure to improve generalization performance in reinforcement learning (RL). However, recent approaches directly use the augmented data to learn the value estimate or regularize the estimation, often ignoring the core essence that the model needs to learn that augmented data indeed represents the same state. In this work, we present **RAIR**: **R**einforcement learning with **A**ugmentation **I**nvariant **R**epresentation that disentangles the representation learning task from the RL task and aims to learn similar latent representations for the original observation and the augmented one. Our approach learns the representation of high-dimensional visual observations in a non-contrastive self-supervised way combined with the standard RL objective. In particular, RAIR gradually pushes the latent representation of an observation closer to the representation produced for the corresponding augmented observations. Thus, our agent is more resilient to the changes in the environment. We evaluate RAIR on all sixteen environments from the RL generalization benchmark Procgen. The experimental results indicate that RAIR outperforms PPO and other data augmentation-based approaches under the standard evaluation protocol.

1 Introduction

Despite the success of modern deep Reinforcement Learning (RL), learning a generalizable policy directly from visual observations remains a foundational concern in modern deep reinforcement learning [Cobbe et al., 2019][Cobbe et al., 2020]. RL algorithms may completely fail to generalize when slight environment changes occur as opposed to the biological agents who can quickly adapt or generalize previous learnings to unseen environmental conditions [Raileanu and Fergus, 2021]. While the use of diverse samples can effectively improve the generalization performance of RL agents, data collection for training real-world control tasks remains costly and notoriously difficult. The online nature of many state-of-the-art RL algorithms exacerbates the issues further. To address this problem, recent approaches have proposed to leverage techniques like data augmentation [Laskin et al., 2020a], regularization [Raileanu et al., 2021], or representation learning [Laskin et al., 2020b]. In this work, we provide a solution from a combined perspective of data augmentation and representation learning.

We observe that current RL algorithms leverage data augmentation in different ways: (i) simply apply augmentation to data and use those augmented data directly for RL objective estimation [Laskin et al., 2020a], (ii) add additional auxiliary losses based on augmented data to regularize the policy and value function [Raileanu et al., 2021][Yarats et al., 2021], and (iii) use unsupervised contrastive learning to learn representation from similar and dissimilar pairs [Laskin et al., 2020b]. For contrastive losses similar data denotes the augmented ones and dissimilar ones refer to the samples from other trajectories [Laskin et al., 2020b]. While contrastive loss-based approaches attempt to learn better representation, naive application of data augmentation often outperforms them

NeurIPS 2023 Workshop on Generalization in Planning (GenPlan 2023).

in challenging tasks where the agent needs to generalize to unseen contexts [Laskin et al., 2020a]. RL algorithms that use augmentation-based regularization aim to directly learn transformation invariant policy and value function, thus sacrificing the explicit representation learning step. In general, learning from state-based features is known to be more sample-efficient than learning from raw pixels. Thus learning latent representations for high-dimensional observations that preserves the semantic meaning can greatly help the control algorithm to achieve sample-efficiency and generalization. Non-contrastive similarity-based approaches have recently achieved success in computer vision (i.e. image classification task) [Bardes et al., 2021][Grill et al., 2020], however, the potential of such an approach is yet to be explored for reinforcement learning.

In this paper, we propose Reinforcement learning with Augmentation Invariant Representation (RAIR) that distinctly considers the representation learning task and the downstream RL task. More specifically, we combine non-contrastive self-supervised embedding learning and RL objective optimization. During the unsupervised representation learning phase, RAIR learns representations that are invariant to the augmented version of the observed state and then goes to the next phase where it optimizes the standard RL objective. By augmentation invariant, we refer to the property that for both the original observation and corresponding transformed observation the encoder will produce a similar latent representation so that the agent can perceive that both are the same. While we validate our approach using Proximal Policy Optimization (PPO) [Schulman et al., 2017] that uses an actor-critic architecture, our method can be used with any RL algorithm (on-policy and off-policy) that uses neural network-based function approximator to represent the policy and value function, and the observation space allows some kind of meaningful transformations of the states. We evaluate our proposed approach on the Procgen benchmark. We empirically show that our proposed approach outperforms standard PPO without any kind of augmentation and is comparable to or better than generalization-specific previous approaches that leverage data augmentation.

2 Related Works

Generalization in Reinforcement Learning (RL) has been extensively addressed in recent years [Farebrother et al., 2018][Packer et al., 2018] [Cobbe et al., 2019]. Several successful methods have emerged to enhance generalization in deep RL. These methods include the application of regularization techniques such as dropout [Igl et al., 2019], batch normalization [Igl et al., 2019][Cobbe et al., 2019][Hu et al., 2021], and data augmentation [Cobbe et al., 2019][Raileanu et al., 2021]. To mitigate observational overfitting, Bertoin and Rachelson [2022] have introduced a feature-swapping regularization technique. Zhang et al. [2020] and Agarwal et al. [2021] have employed bisimulation metrics to analyze the similarity between states, enabling the learning of task-relevant representations. Decoupled policy and value networks have been proposed to avoid overfitting and bolster generalization [Raileanu and Fergus, 2021] [Nafi et al., 2022] [Nafi et al., 2023]. Additionally, policy distillation techniques have been employed to enhance generalization, as demonstrated by Igl et al. [2020] and Lyle et al. [2022]. More recent works have introduced interesting approaches such as the generalist-specialist training framework [Jia et al., 2022] and pre-trained encoder [Yuan et al., 2022].

Data Augmentation is an effective way to learn from limited data and has widely been used in computer vision [Wang et al., 2017][Rebuffi et al., 2021][Goodfellow et al., 2020]. In supervised learning, an augmented version of a sample is assigned the same label and directly used as an independent sample during training [Krizhevsky et al., 2012][Goodfellow et al., 2020]. Self-supervised methods exploit the augmented images to generate pseudo labels for pretext tasks [Chen et al., 2020] [Xie et al., 2020] [Grill et al., 2020][Bardes et al., 2021][He et al., 2020]. Recently, data augmentation has gained significant attention in RL to improve the sample efficiency and generalization [Ma et al., 2022]. The naive approach is to simply use the augmented data as an additional sample for RL optimization [Cobbe et al., 2019] [Lee et al., 2019]. Contrastive loss has been adopted for RL in [Laskin et al., 2020b] that uses crop as the data augmentation technique. Laskin et al. [2020a] presents a comprehensive study on data augmentation for RL and shows that data augmentation can outperform even approaches dedicated to learning better representation. Yarats et al. [2021] uses augmented data to regularize the Q-function in value-based RL algorithms. Following their work, Raileanu et al. [2021] proposes to regularize both the policy and value functions in an actor-critic setting. Raileanu et al. [2021] also presents a way to automatically select the best augmentation technique that achieves better generalization. In our work, we aim to learn augmentation invariant representation for RL using non-contrastive loss.

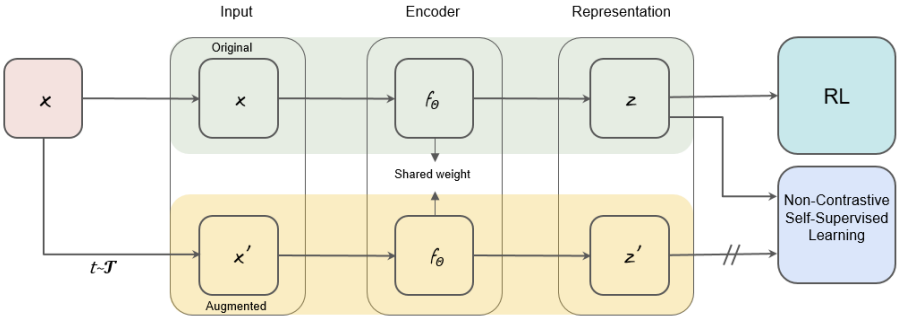

Figure 1: Overview of the proposed non-contrastive RL approach. Unlike other approaches, we directly use the raw sample x and use the augmented one x' only for self-supervised learning.

3 Background

3.1 Contextual Markov Decision Process

Contextual Markov Decision Process (CMDP) extends the general MDP formulation by introducing the dependence on the context. CMDP allows a set of contexts and every context induces a slight variation in the base MDP thus resulting in a number of MDPs that share similar characteristics but vary in terms of initial state distribution and the transition function. Here, we assume a CMDP is defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{C}, \mathcal{T}, r, \mu_C, \mu_S)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{C}$ is the context space, $\mathcal{T}(s'|s, a)$ is the transition function, r is the reward function, μ_C is the context distribution, and μ_S is the context-dependent initial state distribution. Each episode corresponds to a context sampled according to $c \sim \mu_C$. An initial state is sampled according to $s_0 \sim \mu(\cdot|c)$ and the subsequent states within that episode are sampled based on $s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t, c)$. Consider d_π^c as the state distribution that is generated through the execution of the acting policy π in context c. During training, the agent has access to a limited number of contexts. The objective is to learn a generalizable policy π such that the expected return over all possible contexts $\mathcal{G} = \mathbb{E}_{c \sim \mu_C, s \sim d_\pi^c, a \sim \pi(s)}[r(s, a)]$ is maximized.

3.2 Proximal Policy Optimization

Proximal Policy Optimization (PPO) is the widely used policy gradient method [Schulman et al., 2017], and for this work, we use PPO as a baseline and build our approach on top of PPO. While learning from high-dimensional image observation, as the policy and value function approximator, PPO generally leverages a shared neural network. Given that the network is parameterized by θ, PPO optimizes the following objective:

$$J_{PPO}(\theta) = J_\pi(\theta) - \alpha_v L_V(\theta) + \alpha_s S_\pi(\theta) \tag{1}$$

where $J_\pi(\theta)$ is the policy gradient objective, $L_V(\theta)$ is the value loss, $S_\pi(\theta, \phi_\pi)$ is the entropy bonus for exploration, and α_v and α_s are the coefficients for the respective terms.

4 Augmentation Invariant Representation Learning for RL

In this section, we describe our proposed approach RAIR that enables better representation learning for pixel-based observations and consequently improves the generalization ability of the RL agent in zero-shot settings. Our critical insight is that even if two observations visually appear different, if they share the same semantic meaning then the latent representation should be semantically similar. For example, an RGB observation and its grayscale counterpart or a random cropped counterpart denote the same state. Thus, we propose a simple idea to iteratively optimize the state representation encoder so that it can generate similar latent representation in such cases while learning the RL objective. Figure 1 presents an overview of the approach.

Algorithm 1 RAIR: **R**einforcement learning with **A**ugmentation **I**nvariant **R**epresentation

Hyperparameters: total number of updates N, replay buffer size T, minibatch size M, encoder network params ϕ, params for actor-critic heads θ, image transformation t_r with parameters v_i.
for $n = 1, \ldots, N$ **do**
  Collect $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{T}$ using $\pi(\theta)$
  for $j = 1, \ldots, \frac{T}{M}$ **do**
    $\{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{M} \sim \mathcal{D}$
    for $i = 1, \ldots, M$ **do**
      $v_i \sim \mathcal{H}$
      $z_i \leftarrow f_\phi(s_i)$
      $\hat{z}_i \leftarrow f_\phi(t_r(s_i; v_i))$
    end for
    $J(\phi) = \frac{1}{M} \sum_{i=1}^{M} 1 - cos_s(z_i, \hat{z}_i)$
    Train encoder with $J(\phi)$
    $J_{PPO}(\theta, \phi) \leftarrow J_\pi + \alpha_{\text{v}} L_{\text{V}} - \alpha_\pi S_\pi$
    Train full actor-critic network with $J_{PPO}(\theta, \phi)$
  end for
end for

4.1 Augmentation Invariant Representation

We propose to use non-contrastive self-supervised learning for RL to ensure that the representations of behaviorally similar states are not being pushed far apart. Contrastive learning is designed to minimize the distance between anchor-positive pairs (similar) while maximizing the distance between anchor-negative pairs (dissimilar). Thus, similar data points are clustered together, and dissimilar points are separated further. However, the notion of dissimilarity in RL is very different. An observation even being so different in visual appearance may have similar consequences if they are behaviorally similar. Thus, it is hard to distinguish true negative samples without secondary measures. Contrastive learning runs the risk of pushing the behaviorally similar states further in the latent space and disturbs the representation learning task. Therefore, we develop a non-contrastive learning approach to RL that uses only the positive pairs generated through data augmentation. For augmented states, we know for sure that our agent should consider them behaviorally similar.

To this end, we propose to disentangle the representation learning task from the RL task. Our aim is to learn encoders $f_\phi : \mathcal{S} \rightarrow \mathcal{Z}$ that captures representations of states that are augmentation invariant, at the same time suitable for the control task. We train the encoder to minimize the cosine similarity loss which is defined as follows:

$$J(\phi) = 1 - cos_s(z, \hat{z}) \tag{2}$$

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,

$$cos_s(z_i, z_j) = \frac{z_i . z_j}{max(||z_i||_2 . ||z_j||_2, \epsilon)}, \tag{3}$$

where ϵ is a very small number used to avoid division by zero. Other standard similarity metrics such as l_1 distance can also be used in place of cosine similarity. Unlike other algorithms, we directly and solely maximize the agreement between the learned latent representations of similar observations. Our approach also reduces the complexity of learning instance discrimination which has the risk of destabilizing the RL objective.

4.2 Combining Representation Learning with RL

In order to lay out a fully functional reinforcement learning method, we incorporate our augmentation invariant representation learning framework with Proximal Policy Optimization (PPO) which is an actor-critic algorithm. We modify the standard actor-critic algorithm by adding an additional component that calculates the cosine similarity loss and updates the encoder based on that. Algorithm 1 shows all the steps. After the encoder update, RAIR estimates the PPO objective and jointly updates the encoder and the actor-critic parameters. Thus, RAIR iteratively updates the encoder (only) and

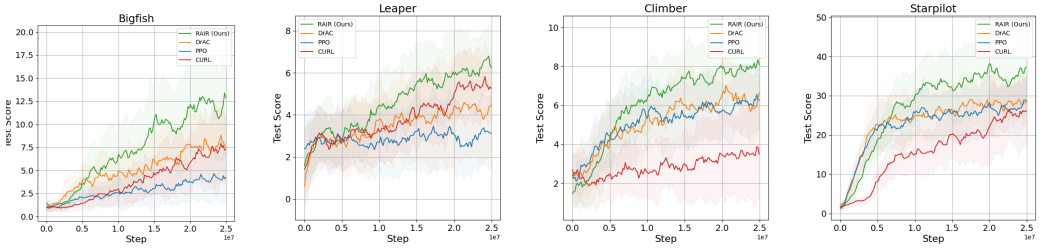

Figure 2: Test performance of RAIR (Ours), DrAC, CURL, and PPO for "crop" data augmentation in four Procgen environments. Mean and std are calculated over 5 trials with different seeds.

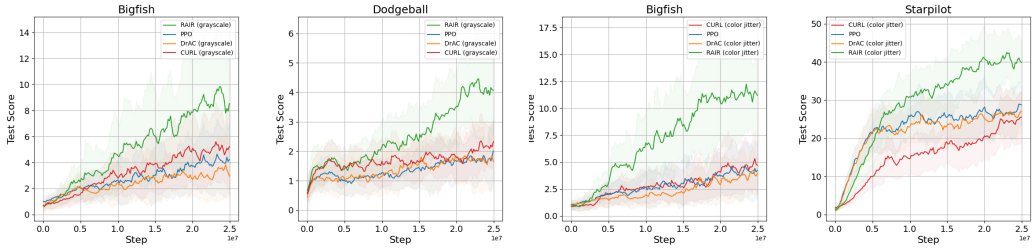

Figure 3: Test performance of RAIR (Ours), DrAC, and PPO for "grayscale" (left two) and "color jitter" (right two) data augmentation in four Procgen environments. Mean and standard deviation are calculated over 5 trials, each with a different seed.

the combined policy and value function (including the encoder) in turn. We observe training with a single combined loss function is often susceptible to instability and high variance. After the training phase, the policy π is used to collect data from the environments. Here, the update sequence is crucial - the encoder update should precede the update for the RL objective. Otherwise, there may be a misalignment between the encoder output and the learned policy. This is because if the encoder update occurs after the policy update, then the mapping learned by the policy head may get disrupted.

5 Experiments

We conduct our experiments on the Procgen benchmark [Cobbe et al., 2020] that offers sixteen procedurally generated environments. Each environment generates levels with diverse dynamics and backgrounds. Each level or episode can be considered as a context as the levels are generated using some seed. Following commonly used standards, we train the agent on 200 contexts or levels and test the agent on the full distribution of contexts. Training on a limited set of levels and testing on the full distribution of possibilities provides the opportunity to evaluate the model's generalization capability. The objective is to achieve better performance in unseen levels or contexts. Figure 2 and Figure 3 present the experimental results for four Procgen environments. RAIR outperforms standard PPO [Schulman et al., 2017], contrastive approach [Laskin et al., 2020b], and generalization-specific approach UCB-DrAC [Raileanu et al., 2021]. We present results for three data augmentation techniques: *crop*, *grayscale*, and *color jitter*.

6 Conclusion

In conclusion, this work introduces RAIR (Reinforcement learning with Augmentation Invariant Representation), a novel approach aimed at enhancing the generalization performance of reinforcement learning (RL) agents through data augmentation. Unlike existing methods that directly use augmented data or contrastive learning, RAIR takes a distinct approach by introducing non-contrastive loss for RL. RAIR learns latent representations of an observation that is invariant to the representation of its corresponding augmented versions. The proposed approach leads to agents that are more robust in adapting to changes in the environment, leading to improved generalization.

References

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

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

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

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

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

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

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=GY6-6sTvGaf`.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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

Jesse Farebrother, Marlos C Machado, and Michael Bowling. Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*, 2018.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. In *International Conference on Artificial Intelligence and Statistics*, pages 829–837. PMLR, 2021.

David Bertoin and Emmanuel Rachelson. Local feature swapping for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Rishabh Agarwal, Marlos C Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. *arXiv preprint arXiv:2101.05265*, 2021.

Nasik Muhammad Nafi, Creighton Glasscock, and William Hsu. Attention-based partial decoupling of policy and value for generalization in reinforcement learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 15–22. IEEE, 2022.

Nasik Muhammad Nafi, Raja Farrukh Ali, and William Hsu. Analyzing the sensitivity to policy-value decoupling in deep reinforcement learning generalization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2625–2627, 2023.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. The impact of non-stationarity on generalisation in deep reinforcement learning. *arXiv preprint arXiv:2006.05826*, 2020.

Clare Lyle, Mark Rowland, Will Dabney, Marta Kwiatkowska, and Yarin Gal. Learning dynamics and generalization in deep reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14560–14581. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/lyle22a.html`.

Zhiwei Jia, Xuanlin Li, Zhan Ling, Shuang Liu, Yiran Wu, and Hao Su. Improving policy optimization with generalist-specialist learning. In *International Conference on Machine Learning*, pages 10104–10119. PMLR, 2022.

Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.

Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8, 2017.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Guozheng Ma, Zhen Wang, Zhecheng Yuan, Xueqian Wang, Bo Yuan, and Dacheng Tao. A comprehensive survey of data augmentation in visual reinforcement learning. *arXiv preprint arXiv:2210.04561*, 2022.

Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*, 2019.