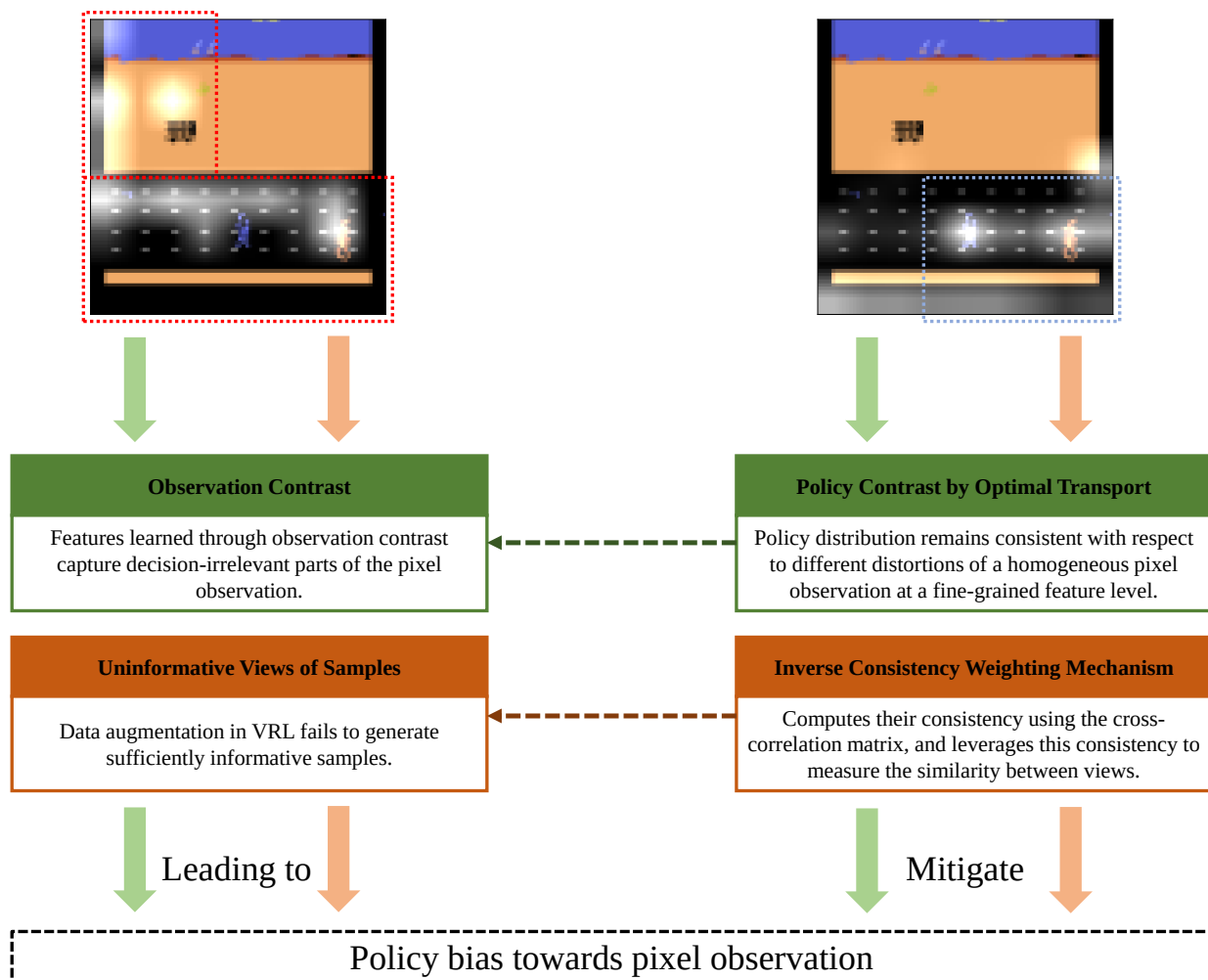


Graphical Abstract

Visual Reinforcement Learning via Sequential Consistency Preserved Policy Contrast from Optimal Transport View

Zehua Zang, Jiangmeng Li, Chuxiong Sun, Rui Wang, Lixiang Liu, Fuchun Sun



Highlights

Visual Reinforcement Learning via Sequential Consistency Preserved Policy Contrast from Optimal Transport View

Zehua Zang, Jiangmeng Li, Chuxiong Sun, Rui Wang, Lixiang Liu, Fuchun Sun

- Our findings reveal that focusing solely on pixel-based features adds little to policy optimization, highlighting the strength of our policy contrast approach.
- Based on our findings, we propose CoCo, a VRL method that leverages policy contrast via the ICW mechanism.
- This paper rigorously evaluates and validates CoCo's underlying mechanism.

Visual Reinforcement Learning via Sequential Consistency Preserved Policy Contrast from Optimal Transport View

Zehua Zang^{a,b,1}, Jiangmeng Li^b, Chuxiong Sun^{b,*}, Rui Wang^b, Lixiang Liu^b and Fuchun Sun^{b,c}

^aUniversity of Chinese Academy of Sciences, Beijing, 101408, China

^bInstitute of Software Chinese Academy of Sciences, Beijing, 100190, China

^cDepartment of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Keywords:

Visual Reinforcement Learning
Contrastive Learning
Optimal Transport

ABSTRACT

Data inefficiency has long posed a significant challenge in the application of visual reinforcement learning methods to complex scenarios. To address this issue, recent researches integrate representation learning mechanisms to extract discriminative features by introducing an auxiliary objective to contrast *pixel observations*. However, our investigations suggest that these representations may not sufficiently capture the essential information necessary for effective decision-making and could potentially impede policy learning. To tackle these limitations, we propose a novel methodology termed *CoCo* (sequential **C**onsistency preserved policy **C**ontrast). Unlike existing approaches, CoCo emphasizes the capture of invariant policy-based discriminative features by performing *policy contrast* across multiple distorted views of observations. Accordingly, we determine that there exists a certain intrinsic heterogeneity between policy and observation, since policy establishes an explicit distribution characteristic rather than a plain tensor. To this end, we propose to model the policy contrast as an optimal transport problem and further perform the alignment of policy distributions during contrastive learning. Subsequently, we introduce an inverse consistency-weighting mechanism designed to accentuate the differences between views while maintaining semantic integrity. Rigorous evaluation across various data efficiency benchmarks consistently demonstrates the effectiveness of CoCo.

1. Introduction

Reinforcement learning has made significant progress in tackling sequential decision-making problems across various domains, including Game AI Bellemare, Naddaf, Veness and Bowling (2013); Samvelyan, Rashid, de Witt, Farquhar, Nardelli, Rudner, Hung, Torr, Foerster and Whiteson (2019) and Robotics Tassa, Doron, Muldal, Erez, Li, de Las Casas, Budden, Abdolmaleki, Merel, Lefrancq, Lillicrap and Riedmiller (2018). However, data inefficiency remains a critical challenge in real-world applications, particularly when dealing with complex pixel observations. To address this, various Visual Reinforcement Learning (VRL) methods Laskin, Srinivas and Abbeel (2020b); Schwarzer, Anand, Goel, Hjelm, Courville and Bachman (2021) have been developed, drawing inspiration from advances in Computer Vision (CV). A pivotal approach in these methods is the introduction of contrastive learning techniques in the observation space Chen, Kornblith, Norouzi and Hinton (2020); Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Pires, Guo, Azar, Piot, Kavukcuoglu, Munos and Valko (2020), referred to as *observation contrast*. Observation contrast enhances the agent's ability to learn a policy by effectively acquiring the agreement between distorted views of a *pixel observation*.

However, we argue that the features, namely pixel-based discriminative features, learned through observation contrast do not align well with the features that are informative for decision-making, i.e., policy-based discriminative features. Accordingly, to support our conclusion, we employ Grad-CAM Selvaraju, Cogswell, Das, Vedantam, Parikh and Batra (2020), a technique that highlights the important regions in the input image influencing the model's decision-making. As illustrated in Figure 1 (a) and (b), features learned through observation contrast capture *decision-irrelevant* parts of the pixel observation, e.g., mountains in BattleZone and road signs in RoadRunner. This phenomenon can be intuitively explained by the nature of contrastive learning: maximizing the mutual information between different distortions

*Corresponding author

✉ zehua2020@iscas.ac.cn (Z. Zang); jiangmeng2019@iscas.ac.cn (J. Li); chuxiong2016@iscas.ac.cn (C. Sun); wangrui@iscas.ac.cn (R. Wang); lixiang@iscas.ac.cn (L. Liu); fcsun@mail.tsinghua.edu.cn (F. Sun)
ORCID(s): 0000-0001-8969-3218 (Z. Zang)

of a common pixel observation at a fine-grained feature level leads the agent to capture the discriminative information only stem from the visual consistency, but there certainly exists a non-trivial gap between such features and the decision-making relevant features. Consequently, plainly performing observation contrast may hinder the policy training process and even lead to training collapse in environments with complex backgrounds. By contrast, an agent is expected to perform decision-making based on policy-based discriminative features, as the optimal policy is influenced solely by the *decision-relevant* parts of the observation. Thus, we propose to explore the *policy contrast* rather than the *observation contrast*.

Considering the policy contrast, we disclose that an intrinsic heterogeneity remains between the policy and the observation, since agents primarily fit a specific policy distribution, e.g., Q-distribution Hosu and Rebedea (2016), Gaussian distribution Haarnoja, Zhou, Abbeel and Levine (2018), and etc., based on the features and further determine the conducted decision by sampling from the policy distribution. In essence, the policy establishes an explicit distribution characteristic rather than a plain tensor, and accordingly, the policy requires a distribution measurement metric instead of the conventional Euclidean distance metric during contrasting. In this regard, the policy distribution of an agent is expected to remain consistent with respect to different distortions of a homogeneous pixel observation at a fine-grained feature level. Therefore, we mandate a shift in typical contrastive learning by framing the policy contrast from the optimal transport perspective Villani et al. (2009); Qiang, Li, Zheng, Su and Xiong (2023); Qiang, Li, Zheng and Su (2021), which provides a powerful approach, e.g., Sinkhorn algorithm Cuturi (2013), for aligning distributions by minimizing the cost of transforming one distribution into another, jointly ensuring stable and scalable convergence to an optimal solution. Compared to observation contrast, features learned by policy contrast, as shown in Figure 1 (c), can significantly reduce the interference of decision-irrelevant parts for agents.

To sufficiently exploit the advantages of policy contrast, it is essential to *enhance* the distinction between different views of the same observation while retaining semantic information, thereby deriving non-trivial representations. This aspect has not been thoroughly explored in VRL. To this end, we introduce an Inverse Consistency Weighting (ICW) mechanism for policy contrast. This approach involves encoding views of a pixel observation into feature, computes their consistency using the cross-correlation matrix, and leverages this consistency to measure the similarity between views. We subsequently assign higher training weights to views with lower consistency, encouraging more effective policy contrast and enhancing the training process. On top of this, we integrate strategy comparison with dynamic model Schwarzer et al. (2021); Yu, Lan, Zeng, Feng, Zhang and Chen (2021) to further improve the agent's ability to learn policy. Empirically, we conduct extensive evaluations on VRL benchmark to demonstrate CoCo's effectiveness. Comprehensive ablation studies confirm that proposed module significantly enhances CoCo's performance. Major **contributions** are as following:

- Our empirical observations reveal a persistent flaw in benchmark VRL methods: focusing solely on pixel-based discriminative features offers limited contributions to policy optimization. This underscores the effectiveness of our proposed policy contrast approach.
- Based on our empirical explorations, we propose CoCo, a novel VRL method that implements policy contrast through ICW mechanism.
- This paper presents rigorous evaluations to validate the effectiveness of the underlying mechanism of CoCo.

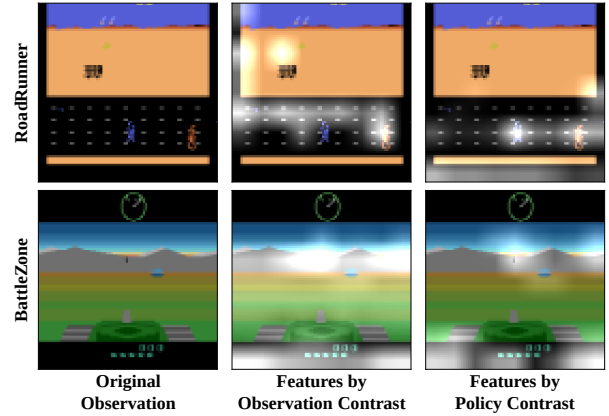


Figure 1: Grad-CAM visualizations on BattleZone and RoadRunner, where brighter areas indicate greater focus by the agent. Origin pixel observations (left). Features learned by observation contrast (middle). Features learned by policy contrast (right).

2. Related Works

2.1. Visual Reinforcement Learning

Traditional VRL methods often suffer from limited data efficiency, especially in environments where learning from pixel observations is required. Beyond approaches that directly optimize policies using reinforcement learning paradigm, such as Kaiser, Babaeizadeh, Milos, Osinski, Campbell, Czechowski, Erhan, Finn, Kozakowski, Levine, Mohiuddin, Sepassi, Tucker and Michalewski (2020); van Hasselt, Hessel and Aslanides (2019); Kielak (2020), a variety of methods have been developed to leverage advancements in CV.

Laskin, Lee, Stooke, Pinto, Abbeel and Srinivas (2020a) and Yarats, Kostrikov and Fergus (2021a) investigate the effectiveness of various data augmentation techniques in CV, such as random cropping, demonstrating that increasing the diversity of training observations can significantly enhance the data efficiency of VRL. Building on this, Raileanu, Goldstein, Yarats, Kostrikov and Fergus (2021) propose an automated augmentation strategy designed to further improve the data efficiency of VRL methods. Additionally, Ma, Zhang, Wang, Li, Wang, Wang, Shen, Wang and Tao (2023) introduces an efficient data augmentation focused on reducing the need for extensive training samples.

Leveraging the strengths of contrastive learning, Laskin et al. (2020b) employs the standard contrastive loss function, InfoNCE van den Oord, Li and Vinyals (2018), to extract features from pixel observations. Similarly, Sun, Qian and Miao (2022) focuses on perturbations within a single task and time step, identifying the most under-learned pixel observations and their augmented versions to improve training efficiency. Additionally, Choi, Lee, Song, Jeon, Sohn and Min (2023) imposes a similarity constraint on both fine-grained dynamics information and global semantic representations, while Liu, Zhang, Li, Zhang and Cheng (2023) integrates a learnable transformation into VRL with minimal adjustments to the architecture and training pipeline. Recognizing the sequential nature of environments, Schwarzer et al. (2021) and ? utilize dynamics models to predict nearby states in latent space. Furthermore, Yu et al. (2021) facilitates trajectory augmentation with unsupervised cycle consistency constraints to enhance feature representation training. Most of these methods adopt an observation contrast approach, compelling the encoder to learn all pixel-based discriminative features to boost VRL efficiency. However, we propose capturing policy-based discriminative features through policy contrast, allowing the agent to minimize the influence of task-irrelevant information and focus on task-relevant details.

2.2. Optimal Transport

In probability theory, the optimal transport problem corresponds to the Wasserstein metric Kantorovich (2006), which compares two distributions under a given metric. This metric has been employed to analyze probabilistic inference algorithms as approximate gradient flows Liu, Zhuo, Cheng, Zhang and Zhu (2019) and to provide robust methods for comparing feature distributions and graphs in computer vision and machine learning Kolouri, Park, Thorpe, Slepcev and Rohde (2017); Kandasamy, Neiswanger, Schneider, Póczos and Xing (2018); Togninalli, Ghisu, Llinares-López, Rieck and Borgwardt (2019). The Gromov-Wasserstein distance Mémoli (2011); Vincent-Cuaz, Flamary, Corneli, Vayer and Courty (2022) extends this concept to compare distributions across different metric spaces, with applications in various domains Demetci, Santorella, Sandstede, Noble and Singh (2020); Fickinger, Cohen, Russell and Amos (2022).

3. Method

3.1. Preliminaries

We examine VRL within the framework of a standard Markov Decision Process (MDP), where an agent interacts with the environment over episodes. An MDP is formally defined as a tuple $(\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)$, where $\mathcal{O}$ denotes the observation space, comprising pixel observations o , $\mathcal{A}$ represents the action space from which actions a are sampled; $\mathcal{P} : \mathcal{O} \times \mathcal{A} \rightarrow \Delta(\mathcal{O})$ is the transition function; $\Delta(\mathcal{O})$ indicates the probability distribution over the next observations; $\mathcal{R} : \mathcal{O} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function; $\gamma \in [0, 1)$ is the discount factor; $d_0 \in \Delta(\mathcal{O})$ and represents the distribution of the initial observation o_0 .

Furthermore, our study focuses on two prominent VRL benchmarks: the Atari Learning Environment (Atari) Bellemare et al. (2013) and the DeepMind Control Suite (DMC) Tassa et al. (2018). Atari is particularly well-suited for illustrating the disparity between pixel-based and policy-based discriminative features due to its complex and dynamic backgrounds. It presents a challenging benchmark in which the agent must take discrete actions based on pixel observations. One of the leading approaches for addressing tasks within the Atari environment is the Deep Q

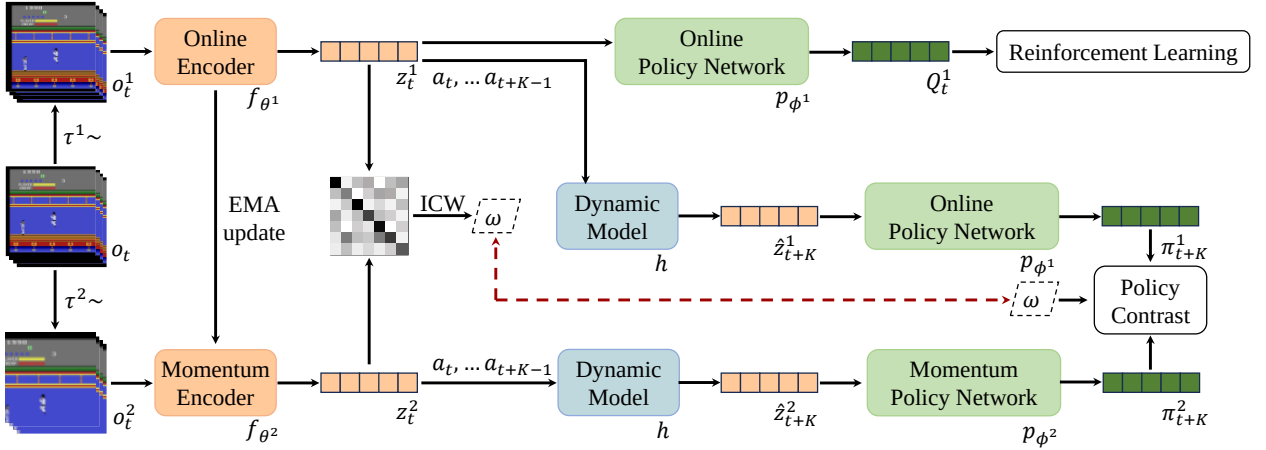


Figure 2: The framework of CoCo. At time step t , an observation o_t is transformed by data augmentations τ^1 and τ^2 into two distinct views o_t^1 and o_t^2 . The online encoder processes the first view o_t^1 to produce the feature z_t^1 , while the momentum encoder processes the second view o_t^2 to generate the feature z_t^2 . These features are then utilized by the dynamic model to predict future features $\hat{z}_{t+K}^1$ and $\hat{z}_{t+K}^2$. The policy network is responsible for generating the policy distributions for each view. Finally, these policy distributions are aligned by Sinkhorn with the ICW mechanism.

Network (DQN) Hosu and Rebedea (2016), which is trained by:

$$\mathcal{L}_\theta^{DQN} = (Q_\theta(o_t, a_t) - (r_t + \gamma \max_a Q_\theta(o_{t+1}, a)))^2, \quad (1)$$

where Q_θ approximates the agent's current Q-function and updates the agent's policy greedily with respect to this Q-function. We used its most widely used variant Rainbow van Hasselt et al. (2019) in our experiments.

Regarding the other benchmark, DMC, we implement SAC Haarnoja et al. (2018). SAC is effective for solving DMC tasks from state observations but struggles to learn effective policies from pixel observations. SAC is an actor-critic method that learns a policy π_ϕ and critics Q_{ϕ_1} and Q_{ϕ_2} . The parameters ϕ_i are learned by minimizing the Bellman error:

$$\mathcal{L}_\phi^{Critic}(\mathcal{B}) = \mathbb{E}_{t \sim \mathcal{B}} [(Q_{\phi_i}(o, a) - (r + \gamma(1 - d)\mathcal{T}))^2], \quad (2)$$

where $t = (o, a, o', r, d)$ is a tuple, $\mathcal{B}$ is the replay buffer, and $\mathcal{T}$ is the target, defined as:

$$\mathcal{T} = \min_{i=1,2} Q_{\phi_i}^*(o', a') - \alpha \log \pi_\phi(a' | o'). \quad (3)$$

While the critic is given by Q_{ϕ_i} , the actor samples actions from policy π_ϕ and is trained by maximizing the expected return of its action as in:

$$\mathcal{L}_\phi^{Actor} = \mathbb{E}_{a \sim \pi} [Q^\pi(o, a) - \alpha \log \pi_\phi(a | o)]. \quad (4)$$

Actions are sampled stochastically from the policy $a_\phi(o, \xi) \sim \tanh(\mu_\psi(o) + \eta_\psi(0) \odot \xi)$ where $\xi \sim \mathcal{N}(0, I)$ is a standard normalized noise vector.

3.2. Policy Contrast with Optimal Transport

Policy contrast is framed from the optimal transport perspective Villani et al. (2009); Qiang et al. (2023, 2021). Let $(\mathcal{A}_1, \pi_1)$ and $(\mathcal{A}_2, \pi_2)$ be two policy spaces, where π_1 and π_2 are policy distributions on $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. Consider a cost function $c : \mathcal{A}_1 \times \mathcal{A}_2 \rightarrow \mathbb{R}^+$ that measures the cost of transporting mass from a point $i \in \mathcal{A}_1$ to a point $j \in \mathcal{A}_2$. The goal of optimal transport is to find a joint distribution γ on $\mathcal{A}_1 \times \mathcal{A}_2$ (a coupling of π_1 and π_2) that minimizes the total transportation cost. The optimal transport problem can be formulated as:

$$\text{OT}_c(\pi_1, \pi_2) = \inf_{\gamma \in \Pi(\pi_1, \pi_2)} \int_{\mathcal{A}_1 \times \mathcal{A}_2} c(i, j) d\gamma(i, j), \quad (5)$$

where $\Pi(\pi_1, \pi_2)$ is the set of all couplings of π_1 and π_2 , i.e., the set of all joint distributions γ on $\mathcal{A}_1 \times \mathcal{A}_2$ such that the marginals are π_1 and π_2 :

$$\gamma(A \times Y) = \pi_1(A) \quad \text{and} \quad \gamma(X \times B) = \pi_2(B) \quad (6)$$

for all measurable sets $A \subset \mathcal{A}_1$ and $B \subset \mathcal{A}_2$.

When the cost function $c(i, j)$ is a metric, say $c(i, j) = d(i, j)^p$ where d is a metric on $\mathcal{A}_1$ and $p \geq 1$, the Wasserstein distance of order p between π_1 and π_2 is defined as:

$$\mathcal{W}(\pi_1, \pi_2) = \left(\inf_{\gamma \in \Pi(\pi_1, \pi_2)} \int_{\mathcal{A}_1 \times \mathcal{A}_2} d(i, j)^p d\gamma(i, j) \right)^{\frac{1}{p}} \quad (7)$$

Overall, given two policy distributions π_1 and π_2 , the primary objective of policy contrast is to minimize the Wasserstein distance between them:

$$\mathcal{J}_{PC} = \min_{\pi_1 \sim \mathcal{A}_1, \pi_2 \sim \mathcal{A}_2} \mathcal{W}(\pi_1, \pi_2) \quad (8)$$

Wasserstein distance is well-defined even when the supports of the distributions do not overlap, as it quantifies the "effort" required to transform one distribution into the other, thus providing enhanced robustness to support mismatch. In contrast, other commonly used distance measures, such as KL divergence, are undefined when the supports of the distributions do not fully overlap, rendering them unsuitable for comparing distributions with disjoint or nearly disjoint supports.

3.3. Implementing Policy Contrast in VRL

Building on the objective function described above, we outline the modifications necessary to adapt it to the experimental scenarios. In the Atari environment, at each time step t , the agent receives a pixel observation o_t from the game. This observation is subsequently transformed through data augmentations $\tau^1 \sim \mathcal{T}^1$ and $\tau^2 \sim \mathcal{T}^2$ into multiple views, denoted as o_t^1 and o_t^2 . The online encoder f_{θ^1} is then utilized to embed the observation o_t^1 into a feature representation, $z_t^1 = f_{\theta^1}(o_t^1)$, and the online policy network p_{ϕ^1} generates the Q-distribution $Q^1(o_t^1) = p_{\phi^1}(z_t^1)$.

Similarly, for o_t^2 , we use a momentum encoder f_{θ^2} and a momentum policy network p_{ϕ^2} to produce the feature representation z_t^2 and the Q-distribution $Q^2(o_t^2) = p_{\phi^2}(z_t^2)$ to implementing contrastive learning. The policy-based discriminative feature is then captured by minimizing the distance between the Q-distributions $Q^1(o_t^1) = p_{\phi^1}(z_t^1)$ and $Q^2(o_t^2) = p_{\phi^2}(z_t^2)$.

We define the basic loss function of policy contrast as:

$$\mathcal{L}_{PC} = \mathcal{W}(Q^1(o_t^1), Q^2(o_t^2)) + \mathcal{W}(Q^1(o_t^2), Q^2(o_t^1)). \quad (9)$$

When applying this approach to DMC, the Q-distribution is replaced by the policy output of the actor network. The momentum encoder and the momentum policy network are updated using the exponential moving average (EMA) Klinker (2011) of the parameters from the online encoder and the online policy network. The update rule for θ_2 and ϕ_2 is as follows:

$$\theta_2 \leftarrow \xi \theta_2 + (1 - \xi) \theta_1, \phi_2 \leftarrow \xi \phi_2 + (1 - \xi) \phi_1, \quad (10)$$

where $\xi \in [0, 1)$ is the coefficient of EMA.

In the DMC environment, we replace the Q-distribution with the policy distribution output by the Actor network for the contrastive learning process. For convenience, we briefly use the symbol π to represent the policy distribution. The resulting loss function for policy contrast is expressed as follows:

$$\mathcal{L}_{PC} = \mathcal{W}(\pi^1(o_t^1), \pi^2(o_t^2)) + \mathcal{W}(\pi^1(o_t^2), \pi^2(o_t^1)), \quad (11)$$

where π_1 represents the output of the online actor, while π_2 denotes the output of the momentum actor.

To further enhance the agent's ability to comprehend sequential patterns in observations and policies, we introduce the dynamic model $h(\cdot, \cdot)$ designed to facilitate the prediction of transition dynamics. Starting with an observation

o_t , its feature is encoded by the online encoder as $z_t = f_{\theta^1}(o_t)$. The online policy network $p_{\phi^1}(\cdot)$ then generates a Q-distribution, from which an action a_t is sampled and converted into a one-hot encoding. The dynamic model h is employed to predict the feature of the next observation, denoted as $\hat{z}_{t+1} = h(z_t, a_t)$.

Following this procedure, given the current observation feature z_t and a sequence of one-hot encoded actions a_{t+K-1} , a sequence of K predictions $\hat{z}_{t+1+K}$ for the future observation features is obtained by iteratively computing:

$$\hat{z}_{t+k+1} = \begin{cases} h(z_{t+k}, a_{t+k}), & k = 0 \\ h(\hat{z}_{t+k}, a_{t+k}), & k = 1, 2, \dots, K-1. \end{cases} \quad (12)$$

The dynamic model is trained through supervised learning, using the predicted transition $(z_t, a_t) \rightarrow \hat{z}_{t+1}$, as follows:

$$\mathcal{L}_{DM} = -\left(\frac{z_{t+k}}{\|z_{t+k}\|_2}\right)^T \left(\frac{\hat{z}_{t+k}}{\|\hat{z}_{t+k}\|_2}\right). \quad (13)$$

3.4. Inverse Consistency Weighting Mechanism

When applying contrastive learning, it is crucial to generate multiple views that differ but still retain the same semantic information. However, in the context of VRL, we observe that the agent struggles to learn a meaningful policy for solving tasks because the data augmentation in VRL fails to generate sufficiently informative samples. To address this issue, we modified our previous data augmentation scheme. Specifically, we slightly increased the augmentation intensity for one of the views while keeping the other view unchanged. Additionally, to further incentivize training on these augmented samples, we propose a Consistent Inverse Weighting (ICW) mechanism. The ICW-based weight is defined as follows:

$$\omega(z_t^1, z_t^2) = \mu \sum_{i=1}^n (-C_{ii} + \frac{1}{2}C_{ii}^2) + \frac{\mu}{2} \sum_{i=1}^n \sum_{j \neq i}^d C_{ij}^2, \quad (14)$$

where $C = (z_t^1) \cdot (z_t^2)^T$, μ is a hyper-parameter, and n represents the dimensionality of $z_t^{(\cdot)}$. The intuition behind ICW is to prioritize the training of under-represented or under-trained observation views, which generally exhibit a higher ICW.

The integrated inverse consistency weighted distribution alignment-based loss function of CoCo is formalized by

$$\mathcal{L}_{CoCo} = \lambda_{ICW} \cdot \omega \cdot \mathcal{L}_{PC}, \quad (15)$$

where λ_{ICW} is the hyper-parameter to scale ICW weight ω . A detailed explanation of the theory behind ICW is provided in Appendix C.

4. Experiments

In this section, we present performance evaluations and conduct extensive ablation studies to demonstrate the effectiveness of CoCo. Implementation details and additional results can be found in Appendix A and Appendix B, respectively.

4.1. Evaluation Benchmarks and Implementations

Our primary goal for CoCo is improve data efficiency while tackling control tasks from pixels. Therefore, we evaluate CoCo on the data efficient VRL benchmark Atari-100K, DMC-100K and DMC-500K.

Atari-100K. The Atari-100K benchmark, introduced by Kaiser et al. (2020), is designed to highlight the challenges posed by complex background information, making it well-suited for demonstrating the gap between pixel-based and policy-based discriminative features. Under this benchmark, the performance of various methods is evaluated across 26 tasks, with 100,000 environment interaction steps of training, equivalent to 400,000 environment frames. Consistent with Kaiser et al. (2020); Yu et al. (2021), we utilize the mean human-normalized score (Mean HNS) and median human-normalized score (Median HNS) as the metrics for performance evaluation. The HNS is calculated as the normalized cumulative reward of an evaluation episode, defined by $HNS = (agent\ score - random\ score) / (human\ score - random\ score)$. The Mean HNS and Median HNS represent the aggregated HNS across the 26 tasks, calculated as the mean and median, respectively. The results are obtained from 10 runs at the end of training and averaged over 20 random seeds.

Table 1

Quantitative evaluations of episode cumulative rewards with 9 state-of-the-art methods on 26 Atari games.

Game	SimPLe	DER	CURL	CCLF	ADAT	DrQ	PSRL	CoIT	SPR	CoCo
Alien	616.9	739.9	558.2	920.0	1029.7	771.2	1030.1	1206.7	801.5	1412.7
Amidar	88.0	188.6	142.1	154.7	147.3	102.8	114.3	182.3	176.3	239.0
Assault	527.2	431.2	600.6	612.4	749.4	452.4	708.3	635.7	571.0	758.6
Asterix	1128.3	470.8	734.5	708.8	864.0	603.5	959.3	709.0	977.8	1036.0
BankHeist	34.2	51.0	131.6	36.0	164.0	168.9	95.8	124.8	380.9	35.2
BattleZone	5184.4	10124.6	14870.0	5775.0	21240.0	12954.0	16688.0	13760.0	16651.0	26080.0
Boxing	9.1	0.2	1.2	7.4	0.4	6.0	35.9	23.6	35.8	41.3
Breakout	16.4	1.9	4.9	2.7	4.5	16.1	17.5	16.1	17.1	21.4
ChopperCommand	1246.9	861.8	1058.5	765.0	1106.0	780.3	1251.2	1338.0	974.8	3860.0
CrazyClimber	62583.6	16185.3	12146.5	7845.0	21240.0	20516.5	42544.0	17538.0	42923.6	51273.1
DemonAttack	208.1	508.0	817.6	1360.9	851.9	1113.4	884.0	846.4	545.2	484.1
Freeway	20.3	27.9	26.7	22.6	29.7	9.8	24.8	29.6	24.4	29.2
Frostbite	254.7	866.8	1181.3	1401.0	1943.2	331.1	776.9	2069.8	1821.5	2579.7
Gopher	771	349.5	669.3	814.7	601.2	636.3	920.3	746.8	715.2	1176.4
Hero	2656.6	6857.0	6279.3	6944.5	7259.2	3736.3	3977.3	7572.8	7019.2	13536.0
Jamesbond	125.3	301.6	471.0	308.8	635.7	236.0	471.4	336.0	365.4	491.0
Kangaroo	323.1	779.3	872.5	650.0	956.9	940.6	1580.0	4116.6	3276.4	1493.0
Krull	4539.9	2851.5	4229.6	3975.0	3502.9	4018.1	4958.3	3426.2	3688.9	4629.1
KungFuMaster	17257.2	14346.1	14307.8	12605.0	19146.0	9111.0	17759.5	9250.0	13192.7	4433.0
MsPacman	1480.0	1204.1	1465.5	1397.5	1075.0	960.5	1597.3	1509.6	1313.2	2442.0
Pong	12.8	-19.3	-16.5	-17.3	-15.1	-8.5	-8.2	1.5	-5.9	-13.8
PrivateEye	58.3	97.8	218.4	100.0	388.0	-13.6	158.0	145.7	124.0	100.0
Qbert	1288.8	1152.9	1042.4	953.8	1578.0	854.4	1290.3	2117.5	669.1	3098.8
RoadRunner	5640.6	9600.0	5661.0	11730.0	12508.0	8895.1	3175.7	11758.5	14220.5	14112.0
Seaquest	683.3	354.1	384.5	550.5	251.6	301.2	734.9	554.0	583.1	416.4
UpNDown	3350.3	2877.4	2955.2	3376.3	3597.8	3180.8	4263.8	4734.2	28138.5	5945.8
Mean Score	0.443	0.285	0.381	0.382	0.470	0.357	0.610	0.543	0.352	0.737
Median Score	0.144	0.161	0.175	0.181	0.206	0.268	0.344	0.352	0.415	0.445

DMC-100K/500K. Following Yarats, Zhang, Kostrikov, Amos, Pineau and Fergus (2021b), we compare CoCo against four state-of-the-art VRL methods across six tasks: FingerSpin, CartpoleSwingup, ReacherEasy, CheetahRun, WalkerWalk, and BallInCupCatch, with training conducted over 100,000 and 500,000 environment interaction steps, respectively. Each task is evaluated five times, and we report the mean and variance of the experimental results.

4.2. Performance on VRL Benchmarks

Table 1 presents the comparisons of CoCo with several state-of-the-art methods, including SimPLe Kaiser et al. (2020), Data-Efficient Rainbow van Hasselt et al. (2019), CURL Laskin et al. (2020b), CCLF Sun et al. (2022), ADAT Zheng, Wang, Sun, Ma, Zhao, Xu, III and Huang (2023), DrQ Yarats et al. (2021a), PSRL Choi et al. (2023), CoIT Liu et al. (2023), and SPR Schwarzer et al. (2021). CoCo achieves a Mean HNS of 0.737 and a Median HNS of 0.445, outperforming the state-of-the-art method SPR by 3.4% on Mean HNS and 3% on Median HNS. CoCo also delivers the best performance on 16 out of the 26 games, surpassing the human expert score on six tasks, which demonstrates that our method achieves superior performance in terms of both data efficiency and asymptotic performance.

Table 2 reports the comparisons of CoCo with state-of-the-art methods, including CURL Laskin et al. (2020b), SPR Schwarzer et al. (2021), DrQ Yarats et al. (2021a), CoDy You, Arenz, Chen and Peters (2022), CCLF Sun et al. (2022),

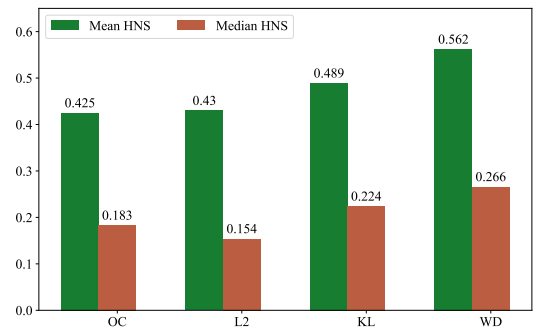


Figure 3: Performance among different distribution alignment implementations.

Table 2

Performance evaluations of episode cumulative rewards on 6 DMC tasks.

100K Step Score	Pixel SAC	CURL	SPR	DrQ	CoDy	CCLF	CCFDM	CoCo (Ours)
FingerSpin	179 ± 66	767 ± 56	868 ± 143	901 ± 42	887 ± 39	944 ± 42	880 ± 142	939 ± 19
CartpoleSwingup	419 ± 40	582 ± 146	799 ± 42	759 ± 92	784 ± 18	799 ± 61	785 ± 87	821 ± 17
ReacherEasy	145 ± 30	538 ± 233	638 ± 269	601 ± 213	624 ± 42	738 ± 99	811 ± 220	809 ± 62
CheetahRun	197 ± 15	299 ± 48	467 ± 36	344 ± 67	323 ± 29	317 ± 38	274 ± 98	477 ± 25
WalkerWalk	42 ± 12	403 ± 24	398 ± 165	612 ± 164	673 ± 94	648 ± 165	634 ± 132	483 ± 67
BallInCupCatch	312 ± 63	769 ± 43	861 ± 233	913 ± 53	948 ± 6	861 ± 233	962 ± 28	934 ± 14
Mean Score	215.7	559.7	671.8	688.3	706.5	726.7	724.3	743.8

500K Step Score	Pixel SAC	CURL	SPR	DrQ	CoDy	CCLF	CCFDM	CoCo (Ours)
FingerSpin	179 ± 166	926 ± 45	924 ± 132	938 ± 103	937 ± 31	974 ± 6	906 ± 152	975 ± 16
CartpoleSwingup	419 ± 40	841 ± 45	870 ± 12	868 ± 10	872 ± 5	869 ± 9	975 ± 38	866 ± 5
ReacherEasy	145 ± 30	929 ± 44	925 ± 79	942 ± 71	957 ± 16	941 ± 48	973 ± 36	938 ± 21
CheetahRun	197 ± 15	518 ± 28	716 ± 47	660 ± 96	656 ± 43	588 ± 22	552 ± 130	737 ± 20
WalkerWalk	42 ± 12	902 ± 43	916 ± 75	921 ± 45	943 ± 17	936 ± 23	929 ± 68	935 ± 14
BallInCupCatch	312 ± 63	959 ± 27	963 ± 8	863 ± 9	970 ± 4	961 ± 9	979 ± 17	968 ± 2
Mean Score	215.7	845.8	885.7	865.3	888.7	878.2	885.7	903.2

Table 3

Performance with and without ICW mechanism

	Mean HNS ↑	Median HNS ↑	Collapse Prob ↓
w/ ICW	0.632	0.333	6.40%
w/o ICW	0.562	0.266	10.16%

Table 4

Top 10% performance of different data augmentation.

	Top 10% Mean HNS	Top 10% Median HNS
NW	0.961	0.498
WS	0.933	0.594

CCFDM Nguyen, Luu, Vu and Yoo (2021). Besides, the performance of pixel SAC is also reported. CoCo outperforms the baselines in two tasks on the DMC-100K and three on the DMC-500K, which strongly support the data efficacy of CoCo, the proposed method, in the domain of VRL. The advanced performance achieved on DMC-100K further demonstrates the excellent data efficiency of CoCo.

4.3. Effectiveness of Policy Contrast

To compare the differences between observation contrast and policy contrast, we conducted two types of experiments. The first experiment involved Grad-CAM analysis. As illustrated in Figure 1, after 100,000 training steps on the Atari games BattleZone and RoadRunner, the features learned through the proposed policy contrast method were more consistent with policy-based discriminative features. Additionally, we implemented policy contrast using three different policy distribution alignment methods: L2 distance (L2), KL divergence (KL), and Wasserstein distance (WD). For the implementation of observation contrast (OC), we employed BYOL Grill et al. (2020) as the observation distribution alignment method. As shown in Figure 3, except for the L2, KL and WD both outperformed OC in both Mean HNS and Median HNS metrics on the Atari-100K benchmark. These results indicate that an appropriate policy contrast method can enable the agent to learn a policy with higher data efficiency.

4.4. Influence of Policy Contrast Implementation

As discussed in Section 3.2, policy contrast can be achieved through various policy distribution alignment methods. However,

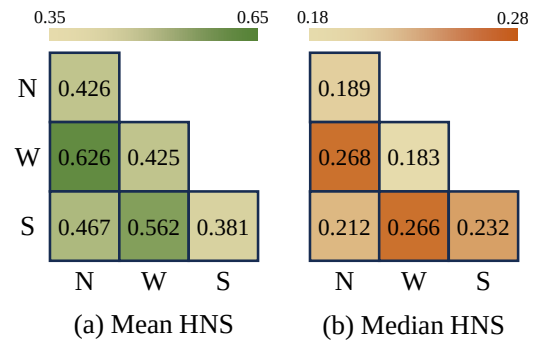


Figure 4: Performance of six types of data augmentation implementations for generating two views of pixel observation.

the choice of alignment method significantly impacts the results. To explore this, our ablation study employed three different policy distribution alignment methods: L2, KL, and WD, as shown in Figure 3. The results demonstrate that policy contrast using Wasserstein distance consistently outperforms other metrics in terms of both Mean HNS and Median HNS. Notably, policy contrast with KL divergence also shows slight improvements over L2 distance. Therefore, the choice of policy distribution alignment method matters, and we implement WD in our proposed method.

4.5. Necessity of ICW Mechanism

To evaluate the extent of improvement that the ICW mechanism brings to policy contrast, we conducted a series of experiments, with the results summarized in Table 3. When policy contrast is implemented with ICW, the Mean HNS and Median HNS show a 7% and 6% improvement, respectively, compared to policy contrast without ICW, confirming the effectiveness of the ICW mechanism. Additionally, we observed a phenomenon where the evaluation results may fall below those of a random policy, which we refer to as training collapse. Without the ICW mechanism, there is a 10.16% probability of training collapse, whereas this probability is reduced to 6.40% when the ICW mechanism is applied. This suggests that the ICW mechanism not only enhances performance but also stabilizes the training process, underscoring its necessity.

4.6. Proper Data Augmentation is Crucial

Data augmentation is crucial for enhancing the agent's performance. To improve its integration with the ICW mechanism, we conducted experiments on different data augmentation schemes that generate views from pixel observations. Specifically, we introduced three levels of augmentation strength. The first scheme, denoted as N(one), involves using the original observation without any augmentation. The second scheme, denoted as W(eak), employs a random shift from DrQ Yarats et al. (2021a), which slightly shifts the observation. The third scheme, denoted as S(trong), uses RandomResizedCrop from BYOL Grill et al. (2020), which applies a significant crop to the observation.

The results of the six combinations of data augmentation strengths are presented in Figure 4. First, applying strong data augmentations to both observation views yields the worst performance in terms of both Mean HNS and Median HNS, which aligns with our previous analysis. When pixel observations are heavily distorted by strong augmentations, much of the policy-based discriminative feature is lost, making it difficult for the views to align effectively, thereby hindering contrastive learning. The results for the NN (no augmentation on either view) and WW (weak augmentation on both views) combinations are consistent with the findings of Chen et al. (2020); Li, Qiang, Zheng, Su and Xiong (2022), which suggest that strong data augmentations are typically necessary for canonical contrastive learning. However, the combination of NW (no augmentation on one view and weak augmentation on the other) achieves the best performance, while WS (weak augmentation on one view and strong augmentation on the other) yields the second-best results.

Although there is a gap in Mean HNS between the NW and WS combinations, their Median HNS values are quite similar. To further investigate, we examined their top 10% performance, as illustrated in Table 4. We observed that the NW combination achieves high performance by securing elevated HNS scores on specific tasks, such as DemonAttack, Hero, JamesBond, and Krull. Conversely, the WS combination attains higher extreme performance due to the effect of more substantial data augmentation. Based on these observations, we opted to use the WS data augmentation strategy for our experiments.

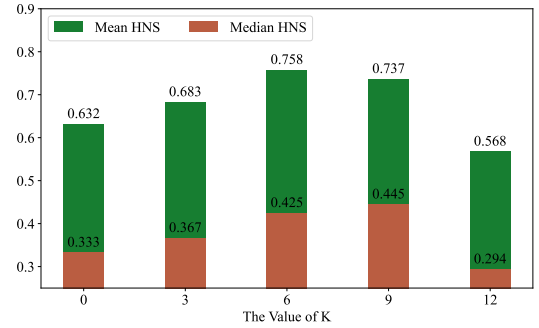


Figure 5: Performance with different dynamic model prediction depths on Atari-100K.

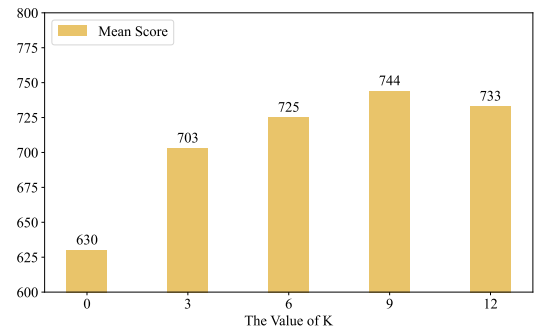


Figure 6: Performance with different dynamic model prediction depths on DMC-100K.

4.7. Impact of Dynamic Prediction Depth

The implementation of a dynamic model enhances the agent’s ability to learn temporal features; however, excessively deep dynamic models can impede the agent’s learning process. To address this, we conducted experiments varying the dynamic prediction depth K within the range $[0, 3, 6, 9, 12]$ to identify the optimal tradeoff. The results on Atari-100K, as shown in Figure 5, reveal that the agent achieves the best performance in terms of Mean HNS with $K = 6$, while the highest performance in terms of Median HNS occurs with $K = 9$. For a comprehensive evaluation across the 26 tasks in the Atari benchmark, we report the results with $K = 9$ as the final performance for our proposed CoCo. However, $K = 6$ is also a viable option for optimizing performance on specific tasks such as CrazyClimber, DemonAttack, and Krull. On DMC-100K, as shown in Figure 6, the agent achieves the best performance in terms of Mean Score with $K = 9$, which is why we set $K = 9$ for the DMC-100K/500K settings. Furthermore, our study indicates that K should be greater than 3 to fully utilize the dynamic model’s capabilities, but it should remain below 12, as performance deteriorates significantly beyond this point. This decline is likely due to the dynamic model’s reduced accuracy in predicting long-term latent observations.

5. Conclusion and Limitations

We introduce CoCo, a VRL method designed to enhance the agent’s ability to learn policy-based discriminative features. To validate the effectiveness of CoCo, we present substantial empirical evidence demonstrating the significance of these features in VRL. CoCo learns representations from observations through policy contrast and the application of ICW. Experimental evaluations reveal that CoCo achieves superior data efficiency and asymptotic performance compared to state-of-the-art VRL methods on benchmarks such as Atari-100K and DMC-100K/500K. However, in environments with very small action spaces, CoCo may face challenges in learning more policy-based discriminative features compared to what could be achieved with observation contrast alone. Furthermore, developing more effective data augmentation techniques for VRL remains a key issue.

A. Implementation Details

A.1. Hyper-Parameters

We provide a complete set of hyper-parameters of CoCo in Table 5 and Table 6.

A.2. Network Architecture

For the representation learning part, CoCo has five: online encoder f_{θ^1} , momentum encoder f_{θ^2} , dynamic model h , online policy network p_{ϕ^1} and momentum policy network p_{ϕ^2} .

A.3. Training Complex

We evaluate the training time of our method and Rainbow DQN for 100K steps on “Alien”, an Atari game. Our method costs 1948s while Rainbow DQN costs 1211s on average, i.e., the time complexity of our method is roughly 1.6 times that of Rainbow DQN. The evaluation is conducted on i7-12700 and RTX 4090. The dimension of the action space of “Alien” is 4, similar to other games. In general, our behavior of adopting Sinkhorn to calculate Wasserstein distance leads to a marginal raising of training time complexity while significantly improving performance.

A.4. Pseudo-Code

Pseudo-code is shown in Algorithm 1.

B. Additional Experimental Results

The human expert and random strategy scores are the same as reported in van Hasselt et al. (2019). We also provide the details of the experiment results of ablation studies. Table 7 is the detailed version of Figure 3. Table 8 is the detailed version of Table 3 and Figure 4.

C. From Information Bottleneck Principle to ICW

Given a set of images X , a pre-set data augmentation to transform them to distorted images Y , and representations are embeded from Y to Z_θ by a neural network with parameter θ . Applied to self-supervised learning, the Information

Table 5
Hyper-parameters on Atari

Hyper-parameter	value
Gray-scaling	True
Frame stack	4
Observation resize	(84, 84)
Augmentation	Random Shift/Crop
Action repeat	4
Traning steps	100K
Max frames per episode	108K
Replay buffer size	100K
Minimum replay size for sampling	2000
Mini-batch size	32
Optimizer	Adam
Optimizer: learning rate	0.0001
Optimizer: β_1	0.9
Optimizer: β_2	0.999
Optimizer: ϵ	0.00015
Max gradient norm	10
Update	Distributional Q
Dueling	True
Support of Q-distribution	51 bins
Discount factor	0.99
Reward clipping	[-1, 1]
Frame stack	
Priority exponent	0.5
Priority correction	(0.4, 1)
Exploration	Noisy nets
Noisy nets parameter	0.5
Evaluation trajectories	100
Replay period every	1 step
Updates per step	2
Multi-step return length	10
Q network: channels	32, 64, 64
Q network: filter size	$8 \times 8, 4 \times 4, 3 \times 3$
Q network: stride	4, 2, 1
Q network: hidden units	256
Target network update period	1
ICM coefficient	0.001
K	9
EMA coefficient	0.99

Table 6
Hyper-parameters on DMC

Hyper-parameter	value
Frame stack	3
Observation rendering	(100, 100)
Observation downsampling	(84, 84)
Augmentation	Random crop, intensity
Replay buffer size	100000
Initial exploration steps	1000
Action repeat	2 finger-spin and walker-walk; 8 cartpole-swingup; 4 otherwise
Evaluation episodes	10
Optimizer	Adam
$(\beta_1, \beta_2) \rightarrow (\theta_f, \xi_h, \xi_b, \omega)$	(0.9, 0.999)
$(\beta_1, \beta_2) \rightarrow$	(0.5, 0.999)
Learning rate (θ_f, ω)	0.0002 cheetah-run 0.001 otherwise
Learning rate (α)	0.0001
Policy batch size (θ_f, ω)	512
Auxiliary batch size (θ_f, ξ_h, ξ_b)	128
Q-function EMA τ	0.01
Critic target update freq	2
Discount factor	0.99
Initial temperature	0.1
Target network update period	1
Target network EMA τ	0.05
K	6

Bottleneck (IB) principle suggests that an ideal representation should be highly informative about the underlying sample while being as invariant as possible to distortions of that sample. This trade-off is captured by the following loss function:

$$IB_\theta = I(Z_\theta, Y) - \beta I(Z_\theta, X), \quad (16)$$

where $I(\cdot, \cdot)$ denotes mutual information, and β is a positive scalar that balances the trade-off between preserving information and achieving invariance to distortions. Utilizing a classical identity for mutual information, we can rewrite the equation as follows:

$$IB_\theta = [H(Z_\theta) - H(Z_\theta|Y)] - \beta[H(Z_\theta) - H(Z_\theta|X)], \quad (17)$$

where $H(\cdot)$ denotes entropy. The conditional entropy $H(Z_\theta|Y)$ cancels out to zero because the function f_θ is deterministic. This implies that the representation Z_θ , when conditioned on the input sample Y , is perfectly determined and thus has zero entropy. Since the overall scaling factor of the loss function is not crucial, we can rearrange Equation 17 as follows:

$$IB_\theta = H(Z_\theta|X) + \frac{1-\beta}{\beta} H(Z_\theta). \quad (18)$$

Table 7

Addition experimental results on Atari.

Game	OC	KL	L2	WD
Alien	774.1	842.0	873.8	1016.8
Amidar	149.9	172.1	164.4	197.4
Assault	688.5	705.2	673.0	707.8
Asterix	746.3	825.0	653.7	806.3
BankHeist	36.0	18.3	19.4	20.0
BattleZone	11086.7	15633.3	10991.1	19280.0
Boxing	17.8	20.7	20.9	30.5
Breakout	13.7	14.4	14.0	17.7
ChopperCommand	1767.6	2703.3	1601.6	2607.0
CrazyClimber	44630.4	56913.3	47256.7	56648.7
DemonAttack	218.2	325.0	242.6	310.5
Freeway	25.8	28.3	27.3	28.1
Frostbite	1295.6	2127.0	736.8	1771.0
Gopher	423.7	632.0	518.8	817.2
Hero	12449.7	11932.3	9855.3	11674.8
Jamesbond	318.8	420.0	294.4	396.7
Kangaroo	237.8	743.3	499.3	795.0
Krull	2841.3	2853.0	3375.1	3516.1
KungFuMaster	1777.6	4333.3	2562.4	3121.7
MsPacman	1595.1	1529.3	1885.0	1857.2
Pong	-16.3	-13.1	-16.7	-15.8
PrivateEye	100.0	100.0	88.0	100.0
Qbert	702.1	741.7	733.2	1478.3
RoadRunner	6756.0	8970.0	8003.1	10125.0
Seaquest	303.2	382.0	318.7	347.5
UpNDown	3569.0	2350.7	3012.0	4498.3
Mean HNS	0.396	0.489	0.430	0.562
Median HNS	0.170	0.224	0.154	0.266

Measuring the entropy of a high-dimensional signal typically demands large amounts of data, often exceeding the size of a single batch. To overcome this challenge, we make the simplifying assumption that the representation Z follows a Gaussian distribution. For a Gaussian distribution, the entropy is simply given by the logarithm of the determinant of its covariance matrix. Consequently, the loss function becomes:

$$IB_{\theta} = \mathbb{E}_X \log |C_{Z_{\theta}|X}| + \frac{1-\beta}{\beta} \log |C_{Z_{\theta}}|, \quad (19)$$

This equation is not precisely the one we optimize in practice. In fact, our loss function is related to the IB loss given in Equation 19 through several simplifications and approximations.

When $\beta \leq 1$, it is evident from Equation 19 that the optimal solution to the IB trade-off is to set the representation to a constant that is independent of the input. This outcome does not yield meaningful representations and can therefore be disregarded. For $\beta > 1$, we observe that the second term in Equation 19 is multiplied by a negative constant. To simplify, we can replace $\frac{1-\beta}{\beta}$ with a new positive constant λ , which is preceded by a negative sign.

In practice, directly optimizing the determinant of covariance matrices does not yield state-of-the-art solutions. Instead, we replace the second term of the loss in Equation 19 with a proxy that involves minimizing the Frobenius norm of the cross-correlation matrix. Assuming that the representations are rescaled to have a norm of 1 along the batch dimension before being input into the loss, this minimization primarily affects the off-diagonal terms of the covariance matrix. The diagonal terms remain fixed at 1 due to the rescaling, while the minimization encourages the off-diagonal terms to approach 0. This surrogate objective, which decorrelates all output units, shares the same global optimum as the original information maximization objective.

Table 8

Addition experimental results on Atari.

Game	NN	NW	NS	WW	WS	SS	Top 10% of NW	Top 10 % of WS
Alien	806.4	992.5	1100.1	759.8	1016.8	731.1	2237.0	1848.0
Amidar	107.4	188.0	155.5	161.9	197.4	133.5	272.7	273.4
Assault	585.5	711.3	649.6	684.6	707.8	629.4	814.4	807.2
Asterix	519.2	918.8	658.0	715.9	806.3	632.5	1428.3	1218.3
BankHeist	21.6	33.4	9.3	34.1	20.0	29.5	108.0	53.7
BattleZone	10554.6	15566.7	13200.0	10973.3	19280.0	7995.0	25400.0	32566.7
Boxing	19.1	28.5	23.9	25.7	30.5	20.1	46.8	48.8
Breakout	13.3	16.8	13.6	15.4	17.7	14.0	24.1	24.3
ChopperCommand	1700.9	2180.9	2282.7	1927.8	2607.0	1486.3	4253.3	5236.7
CrazyClimber	43765.0	57561.6	45678.7	49760.7	56648.7	44876.7	91876.7	82563.3
DemonAttack	304.9	482.4	288.3	244.0	310.5	269.4	1625.3	821.3
Freeway	25.4	28.1	27.4	26.7	28.1	25.9	30.1	30.1
Frostbite	2251.2	1762.0	1123.2	1008.6	1771.0	1280.0	3061.3	2991.7
Gopher	713.7	737.4	515.2	629.9	817.2	402.9	1330.7	1536.0
Hero	10454.0	13258.5	13413.3	9495.0	11674.8	10282.6	13628.0	13680.8
Jamesbond	277.3	495.7	292.0	319.1	396.7	242.8	550.0	550.0
Kangaroo	111.1	810.9	385.3	464.2	795.0	339.8	1220.0	2730.0
Krull	4111.1	4680.4	3514.8	2696.7	3516.1	3187.1	5706.0	5224.7
KungFuMaster	3217.9	3799.8	2513.3	2088.4	3121.7	1605.0	7053.3	5750.0
MsPacman	1408.0	2173.2	1691.7	1589.8	1857.2	1559.8	3267.3	3134.7
Pong	-18.4	-14.3	-17.5	-16.3	-15.8	-17.5	-10.9	-12.1
PrivateEye	89.9	100.0	100.0	78.2	100.0	86.3	100.0	100.0
Qbert	1225.9	1252.9	850.0	854.5	1478.3	671.8	4097.5	4015.8
RoadRunner	3717.8	12782.2	11020.0	6256.0	10125.0	5070.7	20516.7	17360.0
Seaquest	252.5	388.6	336.8	304.7	347.5	314.1	596.0	484.0
UpNDown	4208.7	3888.5	2944.8	2872.1	4498.3	2643.0	6082.7	7173.3
Mean HNS	0.426	0.626	0.467	0.425	0.562	0.381	0.961	0.933
Median HNS	0.189	0.268	0.212	0.183	0.266	0.132	0.498	0.594

Algorithm 1 CoCo Based on Rainbow

Initialize online encoder f_{θ^1} , target encoder f_{θ^2} , dynamic model h , online policy network p_{ϕ^1} and target policy network p_{ϕ^2} .
 Prediction depth K , data augmentation $\mathcal{T}^1, \mathcal{T}^2$, hyper-parameters σ .

while True **do**

 Sample experiments batch $B = (o, a, r, o')$.

 Data augmentation $o^1 = \mathcal{T}^1(o), o^1 = \mathcal{T}^2(o)$.

$z^1 = f_{\theta^1}(o^1), z^2 = f_{\theta^1}(o^2)$.

for i in range(K) **do**

$\hat{z}_k^1 = h(\hat{z}_{k-1}^1, a_{k-1})$

end for

$Q^1 = p_{\phi^1}(z_k^1), Q^2 = p_{\phi^2}(z_k^2)$

 Calculate ICM weight by Equation (7)

 Calculate CoCo loss $\mathcal{L}_{CoCo}$ by Equation (8)

 Calculate RL loss $\mathcal{L}_{rl}$

$\mathcal{L} = \sigma \mathcal{L}_{CoCo} + \mathcal{L}_{rl}$

 update

end while

For consistency with Equation 19, the second term in ICW should ideally be computed using the autocorrelation matrix of one of the online and momentum networks, rather than the cross-correlation matrix between the online and momentum networks. However, in practice, we do not observe a significant difference in performance between these two approaches.

Similarly, it can be easily demonstrated that the first term of Equation 19 shares the same global optimum as the first term of ICW, which maximizes the alignment between representations of pairs of distorted samples.

References

- Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M., 2013. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Intell. Res.* 47, 253–279. URL: <https://doi.org/10.1613/jair.3912>, doi:10.1613/JAIR.3912.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. A simple framework for contrastive learning of visual representations, in: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, PMLR*. pp. 1597–1607. URL: <http://proceedings.mlr.press/v119/chen20j.html>.
- Choi, H., Lee, H., Song, W., Jeon, S., Sohn, K., Min, D., 2023. Local-guided global: Paired similarity representation for visual reinforcement learning, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE*. pp. 15072–15082. URL: <https://doi.org/10.1109/CVPR52729.2023.01447>, doi:10.1109/CVPR52729.2023.01447.
- Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transport, in: *Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300. URL: <https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html>.
- Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., Singh, R., 2020. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv* URL: <https://www.biorxiv.org/content/early/2020/11/11/2020.04.28.066787>, doi:10.1101/2020.04.28.066787, arXiv:<https://www.biorxiv.org/content/early/2020/11/11/2020.04.28.066787.full.pdf>.
- Fickinger, A., Cohen, S., Russell, S., Amos, B., 2022. Cross-domain imitation learning via optimal transport, in: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net*. URL: <https://openreview.net/forum?id=xP3cPq2hQC>.
- Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent - A new approach to self-supervised learning, in: *Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. URL: <https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html>.
- Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: *Dy, J.G., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018, PMLR*. pp. 1856–1865. URL: <http://proceedings.mlr.press/v80/haarnoja18b.html>.
- van Hasselt, H., Hessel, M., Aslanides, J., 2019. When to use parametric models in reinforcement learning?, in: *Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14322–14333. URL: <https://proceedings.neurips.cc/paper/2019/hash/1b742ae215adf18b75449c6e272fd92d-Abstract.html>.
- Hosu, I., Rebede, T., 2016. Playing atari games with deep reinforcement learning and human checkpoint replay. *CoRR* abs/1607.05077. URL: <http://arxiv.org/abs/1607.05077>, arXiv:1607.05077.
- Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Mohiuddin, A., Sepassi, R., Tucker, G., Michalewski, H., 2020. Model based reinforcement learning for atari, in: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net*. URL: <https://openreview.net/forum?id=S1xCPJHtDB>.
- Kandasamy, K., Neiswanger, W., Schneider, J., Póczos, B., Xing, E.P., 2018. Neural architecture search with bayesian optimisation and optimal transport, in: *Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2020–2029. URL: <https://proceedings.neurips.cc/paper/2018/hash/f33ba15effa5c10e873bf3842afb46a6-Abstract.html>.
- Kantorovich, L., 2006. On the translocation of masses. *Journal of Mathematical Sciences* 133. doi:10.1007/s10958-006-0049-2.
- Kielak, K., 2020. Importance of using appropriate baselines for evaluation of data-efficiency in deep reinforcement learning for atari. *CoRR* abs/2003.10181. URL: <https://arxiv.org/abs/2003.10181>, arXiv:2003.10181.
- Klinker, F., 2011. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte* 58, 97–107.
- Kolouri, S., Park, S.R., Thorpe, M., Slepcev, D., Rohde, G.K., 2017. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* 34, 43–59. URL: <https://doi.org/10.1109/MSP.2017.2695801>, doi:10.1109/MSP.2017.2695801.
- Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., Srinivas, A., 2020a. Reinforcement learning with augmented data, in: *Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. URL: <https://proceedings.neurips.cc/paper/2020/hash/e615c82aba461681ade82da2da38004a-Abstract.html>.
- Laskin, M., Srinivas, A., Abbeel, P., 2020b. CURL: contrastive unsupervised representations for reinforcement learning, in: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, PMLR*. pp. 5639–5650. URL: <http://proceedings.mlr.press/v119/laskin20a.html>.
- Li, J., Qiang, W., Zheng, C., Su, B., Xiong, H., 2022. Metaug: Contrastive learning via meta feature augmentation, in: *Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR*. pp. 12964–12978. URL: <https://proceedings.mlr.press/v162/li22r.html>.
- Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., 2019. Understanding and accelerating particle-based variational inference, in: *Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach*,

- California, USA, PMLR. pp. 4082–4092. URL: <http://proceedings.mlr.press/v97/liu19i.html>.
- Liu, S., Zhang, X.S., Li, Y., Zhang, Y., Cheng, J., 2023. On the data-efficiency with contrastive image transformation in reinforcement learning, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net. URL: <https://openreview.net/pdf?id=-nm-rHXi5ga>.
- Ma, G., Zhang, L., Wang, H., Li, L., Wang, Z., Wang, Z., Shen, L., Wang, X., Tao, D., 2023. Learning better with less: Effective augmentation for sample-efficient visual reinforcement learning, in: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/bc26087d3f82e62044fc77752e86737e-Abstract-Conference.html.
- Mémoli, F., 2011. Gromov-wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* 11, 417–487. doi:10.1007/S10208-011-9093-5.
- Nguyen, T.X., Luu, T.M., Vu, T., Yoo, C.D., 2021. Sample-efficient reinforcement learning representation learning with curiosity contrastive forward dynamics model. *CoRR abs/2103.08255*. URL: <https://arxiv.org/abs/2103.08255>, arXiv:2103.08255.
- van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *CoRR abs/1807.03748*. URL: <http://arxiv.org/abs/1807.03748>, arXiv:1807.03748.
- Qiang, W., Li, J., Zheng, C., Su, B., 2021. Auxiliary task guided mean and covariance alignment network for adversarial domain adaptation. *Knowl. Based Syst.* 223, 107066. URL: <https://doi.org/10.1016/j.knosys.2021.107066>, doi:10.1016/J.KNOSYS.2021.107066.
- Qiang, W., Li, J., Zheng, C., Su, B., Xiong, H., 2023. Robust local preserving and global aligning network for adversarial domain adaptation. *IEEE Trans. Knowl. Data Eng.* 35, 3014–3029. URL: <https://doi.org/10.1109/TKDE.2021.3112815>, doi:10.1109/TKDE.2021.3112815.
- Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., Fergus, R., 2021. Automatic data augmentation for generalization in reinforcement learning, in: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 5402–5415. URL: <https://proceedings.neurips.cc/paper/2021/hash/2b38c2df6a49b97f706ec9148ce48d86-Abstract.html>.
- Samvelyan, M., Rashid, T., de Witt, C.S., Farquhar, G., Nardelli, N., Rudner, T.G.J., Hung, C.M., Torr, P.H.S., Foerster, J., Whiteson, S., 2019. The StarCraft Multi-Agent Challenge. *CoRR abs/1902.04043*.
- Schwarzer, M., Anand, A., Goel, R., Hjelm, R.D., Courville, A.C., Bachman, P., 2021. Data-efficient reinforcement learning with self-predictive representations, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net. URL: <https://openreview.net/forum?id=uCQfPZwRaUu>.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359.
- Sun, C., Qian, H., Miao, C., 2022. CCLF: A contrastive-curiosity-driven learning framework for sample-efficient reinforcement learning, in: Raedt, L.D. (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, ijcai.org. pp. 3444–3450. URL: <https://doi.org/10.24963/ijcai.2022/478>, doi:10.24963/IJCAI.2022/478.
- Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T.P., Riedmiller, M.A., 2018. Deepmind control suite. *CoRR abs/1801.00690*. URL: <http://arxiv.org/abs/1801.00690>, arXiv:1801.00690.
- Togninalli, M., Ghisu, M.E., Llinares-López, F., Rieck, B., Borgwardt, K.M., 2019. Wasserstein weisfeiler-lehman graph kernels, in: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 6436–6446. URL: <https://proceedings.neurips.cc/paper/2019/hash/73fed7fd472e502d8908794430511f4d-Abstract.html>.
- Villani, C., et al., 2009. Optimal transport: old and new. volume 338. Springer.
- Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., Courty, N., 2022. Semi-relaxed gromov-wasserstein divergence and applications on graphs, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net. URL: <https://openreview.net/forum?id=RShaMexjc-x>.
- Yarats, D., Kostrikov, I., Fergus, R., 2021a. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net. URL: <https://openreview.net/forum?id=GY6-6sTvGaf>.
- Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., Fergus, R., 2021b. Improving sample efficiency in model-free reinforcement learning from images, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press. pp. 10674–10681. URL: <https://doi.org/10.1609/aaai.v35i12.17276>, doi:10.1609/AAAI.V35I12.17276.
- You, B., Arenz, O., Chen, Y., Peters, J., 2022. Integrating contrastive learning with dynamic models for reinforcement learning from images. *CoRR abs/2203.01810*. URL: <https://doi.org/10.48550/arXiv.2203.01810>, doi:10.48550/ARXIV.2203.01810, arXiv:2203.01810.
- Yu, T., Lan, C., Zeng, W., Feng, M., Zhang, Z., Chen, Z., 2021. Playvirtual: Augmenting cycle-consistent virtual trajectories for reinforcement learning, in: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 5276–5289. URL: <https://proceedings.neurips.cc/paper/2021/hash/2a38a4a9316c49e5a833517c45d31070-Abstract.html>.
- Zheng, R., Wang, X., Sun, Y., Ma, S., Zhao, J., Xu, H., III, H.D., Huang, F., 2023. TACO: temporal latent action-driven contrastive loss for visual reinforcement learning. *CoRR abs/2306.13229*. URL: <https://doi.org/10.48550/arXiv.2306.13229>, doi:10.48550/ARXIV.2306.13229, arXiv:2306.13229.