Mastering Visual Reinforcement Learning via Positive Unlabeled Policy-Guided Contrast

Zehua Zang[1,2[0000-0001-8969-3218]], Qirui Ji[1,2[0009-0006-5982-9739]], Rui Wang[2[0000-0001-5369-9116]], Kai Li[1,2[0000-0001-8452-620X]], Lixiang Liu[2[0000-0002-3890-1758]], Fuchun Sun[2,3[0000-0003-3546-6305]]

[1] University of Chinese Academy of Sciences, Beijing, 101408, China
[2] Institute of Software Chinese Academy of Sciences, Beijing, 100190, China
[3] Tsinghua University, Beijing, 100084, China
wangrui@iscas.ac.cn

Abstract. Reinforcement learning (RL) has garnered considerable attention in recent years. A fundamental yet challenging problem in this paradigm is the effective perception of high-dimensional environmental information, leading to the emergence of Visual Reinforcement Learning (VRL). This sub-field focuses on learning representations from pixel observations to optimize decision-making policies. In this article, we provide a comprehensive analysis of existing benchmark frameworks and highlight a persistent paradox that challenges current approaches: during different training phases, the exploration of visual semantic information can both enhance and hinder the quality of learned representations. Furthermore, we reveal that the issue of over-redundancy commonly limits the sample efficiency of baseline methods. To address these limitations, we propose a novel plug-and-play approach for VRL. Our method employs a positive unlabeled policy-guided contrastive learning framework to jointly capture anti-redundant and policy-relevant pixel semantics during training. We validate the effectiveness of our approach through extensive benchmarking experiments, demonstrating its superior performance over existing methods in pixel environments.

Keywords: Visual Reinforcement Learning, Self-Supervised Learning, Contrastive Learning

1 Introduction

In recent years, RL has achieved significant progress across various domains, including Game AI [1,2,3,4,5,6] and Robotics [7,8,9]. These successes largely rely on well-defined state design, which are often impractical in real-world applications. Consequently, VRL has emerged as a paradigm aimed at directly learning state representations from high-dimensional observations, such as images, and videos. However, VRL encounters significant sample inefficiency challenges, especially in complex tasks with high-dimensional pixel observations. Drawing from advances in data augmentation and self-supervised learning in image and video processing, VRL aims to enhance visual semantic extraction for improved efficiency. This goal is primarily pursued through data augmentation [10] and self-supervised learning approaches [11].

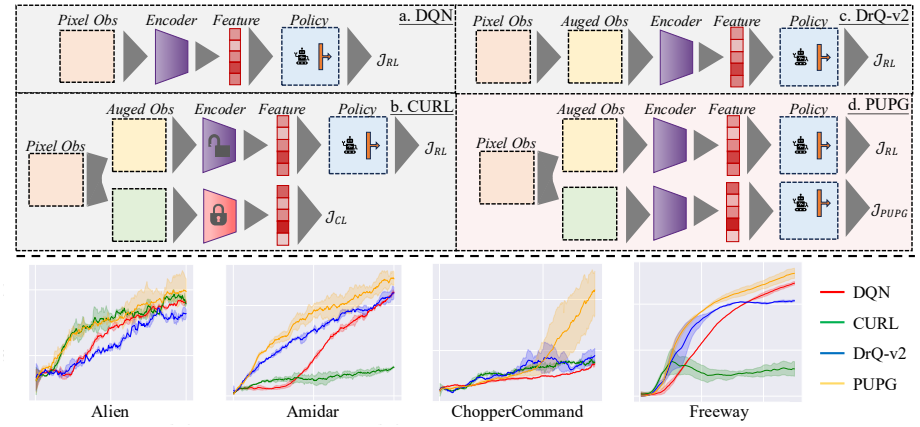

Fig. 1. Simple diagram of DQN, CURL, DrQ-v2 and PUPG, and the reward curves.

To investigate the challenges in VRL, we present frameworks illustrating the learning paradigms of benchmark methods, including DQN [12], CURL [11], DrQ-v2 [10], and our proposed PUPG. We also conduct experimental analyses on reward curves to identify their limitations, as shown in **Fig. 1**. Specifically, RL methods such as DQN and SAC adopt the conventional reward-based training paradigm, where the long-standing issue of limited sample efficiency significantly degrades performance. And the RL encoder struggles to capture rich visual semantic information. The reliance on a reward-dependent training objective constrains the encoder's ability to extract meaningful semantic representations from pixel observations, making it susceptible to becoming trapped in a local optimum. This phenomenon demonstrates that the PUPG consistently outperforms convention-al RL benchmarks, substantiating the aforementioned claims. CURL employs a contrastive learning paradigm to guide the encoder in capturing visual semantic information, thereby improving sample efficiency during training. However, the visual representations learned by the encoder are largely irrelevant to the policy network because the contrastive objective is decoupled from the RL optimization process. This misalignment interferes with the overall training of the RL model, ultimately degrading its final performance. Furthermore, CURL relies on a large number of negative samples to perform effective self-supervision, which reduces training efficiency. DrQ-v2 mitigates sample inefficiency to some extent by integrating data augmentation into the reward-based training paradigm. However, it does not fully address the inherent limitations of reward-based training—specifically, the failure to capture comprehensive visual semantic information. Specifically, capturing rich visual semantic information can enhance both the sample efficiency and the performance potential of RL models. However, overemphasizing the acquisition of visual semantics can interfere with the RL model's training process, preventing it from achieving the global optimum required for effective policy optimization.

Exploring visual semantic information in RL is a double-edged sword, it simultaneously enhances sample efficiency and model performance while increasing the risk of training interference and representation redundancy. To address these challenges, we

propose **Positive Unlabeled Policy-Guided contrast (PUPG)** for VRL, which aims to learn anti-redundant visual semantic information through a balanced trade-off approach. PUPG introduces a positive unlabeled contrastive learning paradigm that mitigates the benchmark methods' reliance on large sets of negative samples while increasing the information entropy of the representations learned by the encoder. This design effectively alleviates the negative impact of VRL's limited sample efficiency. The anti-redundancy property of the learned representations ensures a higher optimization potential for the model. Additionally, by leveraging policy optimization as a guiding mechanism, PUPG regulates the learning process to prevent the acquisition of trivial visual semantics, thereby reducing the risk of the model becoming trapped in undesired local optima. Empirically, we conduct extensive experiments on the widely-used benchmark. Additionally, ablation studies validate the effectiveness of each component within PUPG. The primary contributions of this work are summarized as follows:

1. We disclose the long-standing issues challenging benchmark VRL methods: the duality of exploring the visual semantic information towards the training of the RL model and the over-redundancy problem generally existing in benchmarks.
2. We propose PUPG for VRL, orthogonal to existing benchmark methods, which learns anti-redundant and policy-optimization-relevant visual semantic information for the training of the RL model.
3. We conduct extensive evaluations on various VRL experimental settings to empirically prove the effectiveness of each part of our proposed method.

2 Related Works

2.1 Self-Supervised Contrastive Learning in Computer Vision

Self-supervised contrastive learning aims to learn rich representations of high dimensional unlabeled data by bringing positive samples closer while separating negative samples from each other [13,14,19]. SimCLR [15] presents a framework with a large batch size. MoCo [16] use a dynamic dictionary with a queue to avoid using large batch size. BYOL [17] use the momentum encoder to produce representations of the targets to stabilize the bootstrap step. BarlowTwins [18] measure the cross-correlation matrix between the outputs.

2.2 Sample Efficiency in VRL

Early works [12,27] regard VRL as an end-to-end optimization problem. Meanwhile, reconstruction-based methods [20] combine self-supervised learning and RL by implementing an auto-encoder as an auxiliary task. Works like [10] only apply data augmentation to improve sample efficiency. CURL [11] learns the representations from pixel observations with InfoNCE loss, SPR [21] implements a negative-free self-supervised contrastive learning approach to imposing the similarity constraint between self-predictive and target representations.

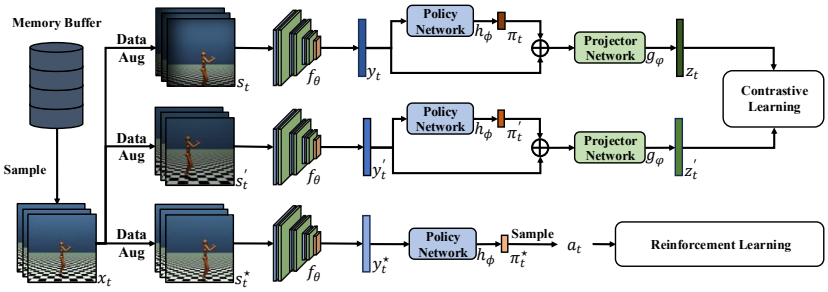

Fig. 2. Training process of PUPG. First, the pixel observation x_t are transformed into multiple augmented views $s_t, s_t', s_t^\star$ by data augmentation. Each view is processed by an encoder f_θ, resulting in feature representations y_t, y_t', $y_t^\star$. During contrastive learning, the policy network h_ϕ generates policy denoted by π_t and π_t', using the features y_t and y_t'. These policy logits are then combined with y_t and y_t' and passed through a projector network g_φ to produce projections z_t and z_t'. During reinforcement learning, given $y_t^\star$, an action is sampled from the policy network h_ϕ, and optimized with a reinforcement learning loss.

3 Preliminaries

VRL is formulated as an Markov Decision Process which is described as a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}: \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ is the transition function, $\Delta(\mathcal{X})$ indicates the next state, $\mathcal{R}: \mathcal{X} \times \mathcal{A} \to [0,1]$ is the reward function, $\gamma \in [0,1)$is a discount factor, and $d_0 \in \Delta(\mathcal{X})$ is the distribution of the initial state x_0. The goal is finding a policy $\pi: \mathcal{X} \to \Delta(\mathcal{A})$ that maximizes the expected cumulative discount rewards $E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$, where $x_0 \sim d_0$.

4 Method

PUPG can be integrated as an auxiliary task alongside any RL algorithm. This section elaborates on three key components of PUPG: data augmentation, policy-guided representation learning, and positive unlabeled contrastive learning. The comprehensive framework of PUPG is illustrated in **Fig. 2**.

Data Augmentation. Semantic information is typically more structured and concentrated in CV datasets, whereas it is more dispersed and varied in VRL environments. Although strong random cropping retains sufficient semantics for representation learning, it may remove critical information needed for policy learning in VRL. To balance augmentation strength and semantic preservation, PUPG adopts a weak random resized crop, ensuring effective augmentation without compromising learning performance.

Policy-Guided Representation. The learned representation should guide the agent toward a better policy, not just encode observations. Relying solely on observation representation may assist early training but often leads to suboptimal policies. To address this, we propose a policy-guided representation that captures task-relevant semantic information for better supporting policy optimization, as:

$$Z = \big(f(s), \pi(s)\big) = \big(f(s;\theta), h(f(s;\theta);\phi)\big) = y \oplus \pi \tag{1}$$

where f is the encoder with its parametersθ, h is the policy network with its parameters ϕ and, $\oplus$ stands for concatenation.

Positive Unlabeled Contrastive Learning. Defining negative views relative to an anchor in pixel observations is inherently challenging. Thus, we adopt the maximum entropy principle, which ensures that the learned representations maintain a balanced and unbiased structure while maximizing their informativeness. Entropy is defined initially on probability distributions, i.e., $H(z) \triangleq -\int p(z)\, log\, p\,(z) dz$, for continuous random variables. However, it is challenging to estimate the true distributions $p(x)$ of representation, from a finite set of high dimensional vectors $Z = [z^1, z^2, \dots, z^m] \in R^{d\times m}$ and encoding the data lossless in infeasible. Instead, we exploit the coding length in lossy data coding as a computationally tractable surrogate for the entropy of continuous random variables. Given a set of samples Z, the minimal number of bits encode Z subject to a distortion ϵ is given by the following coding length function:

$$\mathcal{L} \triangleq (\frac{m+d}{2})\, \text{logdet}\, (I_m + \frac{d}{m\epsilon^2})Z^T Z \tag{2}$$

where I_m denotes the identity matrix with dimension m, and ϵ is the upper bound of the expected decoding error between $z \in Z$ and the decoded $\hat{z}, i.e., E\big||z-\hat{z}|\big|_2 \le \epsilon$.

The computation of the log-determinant of a high dimensional matrix in **Equation (2)** suffers from high expense and may cause numerically unstable results for an ill-conditioned matrix. We rewrite **Equation (2)** as $L = \mu \log\det(I_m + \lambda Z^T Z')$, where $\mu = \frac{m+d}{2}$ and $\lambda = \frac{d}{m\epsilon^2}$, Z' is another view of the sample Z. Utilizing the identical **Equation (2)**, then it becomes $L = \text{Tr}(\mu \log(I_m + \lambda Z^T Z'))$, where Tr stands for the trace of the matrix. With Taylor series expansion applied to it, we obtain the 1^{st}-order, 2^{nd}-order, and 3^{rd}-order expansion:

$$\mathcal{L}^{1^{\text{st}}}_{\text{PUPG}} = \text{Tr}(\mu\lambda Z^T Z') \tag{3}$$

$$\mathcal{L}^{2^{\text{nd}}}_{\text{PUPG}} = \text{Tr}\left(\mu\lambda Z^T Z' - \frac{\mu}{2}(\lambda Z^T Z')^2\right) \tag{4}$$

$$\mathcal{L}^{3^{\text{rd}}}_{\text{PUPG}} = \text{Tr}(\mu\lambda Z^T Z' - \frac{\mu}{2}(\lambda Z^T Z')^2 + \frac{\mu}{3}(\lambda Z^T Z')^3) \tag{5}$$

To balance computational complexity and performance, we implement our policy-guided representation using the 2^{nd}-order expansion in **Equation (4)** with experimental validation provided in **Section 6.7**. And our positive unlabeled policy-guided contrastive loss can be formulated as:

$$\mathcal{L}_{\text{PUPG}} = \mu \underbrace{\sum_{i=1}^{d}\left(-C_{ii} + \frac{1}{2}C_{ii}^2\right)}_{\text{Policy-Guided Consistency}} + \frac{\mu}{2} \underbrace{\sum_{i=1}^{d}\sum_{j\neq i}^{d} C_{ij}^2}_{\text{Policy-Guided Informativeness}} \tag{6}$$

where $C = \lambda Z' Z^T$. Notably, our policy-guided contrastive loss can be jointly trained with any other VRL algorithm.

5 Theoretical Comparison to Existing Contrastive Losses

Relation to InfoNCE. InfoNCE is a widely used contrastive learning loss. And its formulation is given by:

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{m}\left(c_{i,i}/\tau\right) + \sum_{i=1}^{m}\log\sum_{j\neq i}^{m}\left(\exp\left(c_{i,j}/\tau\right) + \exp\left(c_{i,i}/\tau\right)\right) \tag{7}$$

where τ is the temperature parameter. Although **Equation (4)** and the InfoNCE formulation differ in their explicit forms, they exhibit similar effects on the learning process.

Relation to Barlow Twins. Barlow Twins aligns representations by encouraging the cross-correlation matrix between twin embeddings to approximate the identity matrix, which is defined as follows:

$$\mathcal{L}_{\text{Barlow}} = \sum_{i=1}^{d}(1 - C_{ii})^2 + \lambda_{\text{Barlow}}\sum_{i=1}^{d}\sum_{j\neq i}^{d} {C_{ij}}^2, \tag{8}$$

where C denotes the feature-wise cross-correlation matrix and λ_{Barlow} is a positive regularization coefficient. These formulations can be viewed as a 2^{nd}-order approximation of the objective in **Equation (4)**.

Relation to BYOL. BYOL encourages consistency between the representations of differently augmented views of the same input, which is defined as follows:

$$\mathcal{L}_{\text{BYOL}} = ||q - z||_2^2 = 2 - 2\cdot\frac{\langle q, z\rangle}{||q||_2||z||_2}, \tag{9}$$

where q and z are the embeddings of two views of the same observation. Since μ and λ is a constant and can be absorbed by adjusting the learning rate during optimization in **Equation (3)**, BYOL can be viewed as the 1^{st}-order expansion of **Equation (3)**.

6 Experiments

This section empirically evaluates PUPG on an extensive set of VRL environments.

6.1 Performance on Sample Efficiency

To demonstrate the sample efficiency of PUPG, we evaluate it on Atari-100K [22] and DMC-100K/500K [23]. For Atari-100K, we select DQN [12], CURL [11], DrQ-v2 [10], ADAT [24], PSRL [25], and CoIT [26] to compare. 错误!未找到引用源。 demonstrate that PUPG achieves SOTA performance in 12 games.

Table 1. Quantitative evaluations of episodic cumulative rewards on 26 Atari games.

Atari-100K	DQN	CURL	DrQ-v2	ADAT	PSRL	CoIT	PUPG
Alien	739.9	558.2	771.2	1029.7	1030.1	**1206.7**	1168.2
Amidar	188.6	142.1	102.8	147.3	114.3	182.3	**195.1**
Assault	431.2	600.6	452.4	**749.4**	708.3	635.7	727.5
Asterix	470.8	734.5	603.5	864.0	959.3	709.0	**1036.3**
BankHeist	51.0	131.6	**168.9**	164.0	95.8	124.8	50.7
BattleZone	10124.6	14870.0	12954.0	**21240.0**	16688.0	13760.0	17900.0
Boxing	0.2	1.2	6.0	0.4	35.9	23.6	**35.6**
Breakout	1.9	4.9	16.1	4.5	17.5	16.1	**18.9**
ChopperCommand	861.8	1058.5	780.3	1106.0	1251.2	1338.0	**2406.7**
CrazyClimber	16185.3	12146.5	20516.5	21240.0	42544.0	17538.0	**61714.0**
DemonAttack	508.0	817.6	**1113.4**	851.9	884.0	846.4	396.6
Freeway	27.9	26.7	9.8	29.7	24.8	29.6	**30.5**
Frostbite	866.8	1181.3	331.1	1943.2	776.9	2069.8	**2198.2**
Gopher	349.5	669.3	636.3	601.2	920.3	746.8	**1027.9**
Hero	6857.0	6279.3	3736.3	7259.2	3977.3	7572.8	**13503.4**
Jamesbond	301.6	471.0	236.0	**635.7**	471.4	336.0	399.0
Kangaroo	779.3	872.5	940.6	956.9	1580.0	**4116.6**	524.7
Krull	2851.5	4229.6	4018.1	3502.9	**4958.3**	3426.2	3968.8
KungFuMaster	14346.1	14307.8	9111.0	**19146.0**	17759.5	9250.0	5548.0
MsPacman	1204.1	1465.5	960.5	1075.0	1597.3	1509.6	**2500.4**
Pong	-19.3	-16.5	-8.5	-15.1	-8.2	**1.5**	-14.3
PrivateEye	97.8	218.4	-13.6	**388.0**	158.0	145.7	100.0
Qbert	1152.9	1042.4	854.4	1578.0	1290.3	**2117.5**	993.0
RoadRunner	9600.0	5661.0	8895.1	**12508.0**	3175.7	11758.5	11676.7
Seaquest	354.1	384.5	301.2	251.6	**734.9**	554.0	428.4
Mean HNS	0.285	0.381	0.357	0.470	0.610	0.543	**0.637**

For DMC-100K/500K, we select SAC [27], CURL [11], CCFDM [28], SPR [21], CCLF [29], CoDy [30] and MLR [31] as baselines. **Table 2** show that our proposed method consistently achieves the highest episodic cumulative scores across most DMC tasks. PUPG outperforms all baseline methods in these benchmarks, highlighting the robustness and effectiveness of PUPG.

Table 2. Quantitative evaluations of episodic cumulative rewards on 6 DMC tasks.

DMC-100K	SAC	CURL	CCFDM	SPR	CCLF	CoDy	MLR	PUPG
Finger, Spin	179±166	767±56	880±142	868±143	**944±42**	887±39	907±58	876±133

Cartpole, SwingUp	419±40	582±146	785±87	799±42	799±61	784±18	806±48	**840±36**
Reacher, Easy	145±30	538±233	811±220	638±269	738±99	624±42	866±103	**890±102**
Cheetah, Run	197±15	299±48	274±98	467±36	317±38	323±29	482±38	**499±101**
Walker, Walk	42±12	403±24	634±132	398±165	648±110	673±94	643±114	**811±99**
BallInCup, Catch	312±63	769±43	**962±28**	861±233	914±20	948±6	933±16	901±65

DMC-500K	SAC	CURL	CCFDM	SPR	CCLF	CoDy	MLR	PUPG
Finger, Spin	179±166	926±45	906±152	924±132	974±6	937±41	973±31	**980±38**
Cartpole, SwingUp	419±40	841±45	**975±38**	870±12	869±9	869±4	872±5	840±25
Reacher, Easy	145±30	929±44	973±36	925±79	941±48	957±16	957±41	**977±48**
Cheetah, Run	197±15	518±28	552±130	716±47	588±22	656±43	674±37	**743±78**
Walker, Walk	42±12	902±43	929±68	916±75	936±23	943±17	939±10	**948±82**
BallInCup, Catch	312±63	959±27	979±17	963±8	961±9	970±4	964±14	**982±58**

6.2 Advantages in Long-Term Training

To further evaluate the long-term training benefits of PUPG, we selected 10 Atari-100K games where PUPG does not achieve SOTA performance and trained for 2 million steps. As shown in **Fig. 3**, PUPG outperform the baselines.

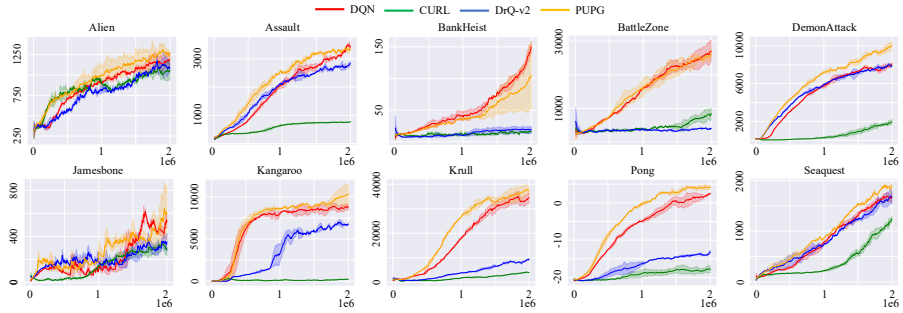

Fig. 3. Performance on 10 Atari games for 2M steps.

6.3 Comparison among Different Negative-Free Contrastive Losses

As shown in **Fig. 4**, PUPG achieves the highest overall performance among the compared methods, with Barlow Twins following closely. Although BYOL does not match the performance of PUPG or Barlow Twins, it still surpasses Rainbow DQN.

Fig. 4. Performance on 3 Atari Games for 100K steps.

6.4 Ablation: Contributions of Each Module

To investigate the properties of PUPG, we set up a series of ablation studies, where we evaluate the contrastive learning with observation representation (PUPG-OC) and contrastive learning with policy (PUPG-PC). As shown in **Fig. 5**, the positive unlabeled contrastive loss is not worse than that in CURL. And strengthening the strength of data augmentation contributes to semantic information learning.

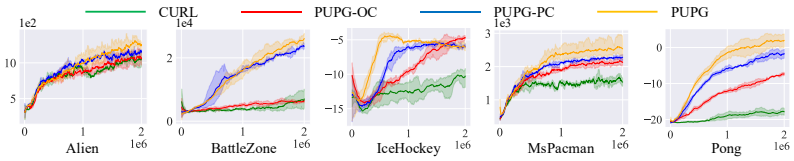

Fig. 5. Performance on 5 Atari games.

6.5 Ablation: Influence of Different Strength of Data Augmentation

As shown in **Section 4**, data augmentation strength—especially the random resized crop—significantly impacts VRL. The crop scale parameter s controls this by setting how much of the original image is retained before resizing. We tested crop scales s = [1.0, 0.85, 0.7, 0.4, 0.1] on three Atari games. As shown in **Fig. 6**, moderate cropping (s = 0.85) improves performance over no augmentation (s = 1.0), but aggressive cropping (s = 0.1) harms performance, likely by removing critical visual information.

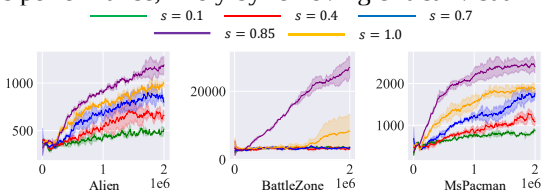

Fig. 6. Performance of different strength of random resized crop on 3 Atari games.

6.6 Visualization of Grad-CAM

As shown in **Fig. 7**, the visualizations reveal that in RL environments, task-relevant information is often confined to a limited portion of the pixel observations. Both CURL and PUPG, which explore visual semantic information, learn representations capturing more task-relevant details (e.g., brick status in Breakout) compared to DQN. By leveraging the informativeness term, PUPG further outperforms benchmark methods, learning more informative and task-aligned representations.

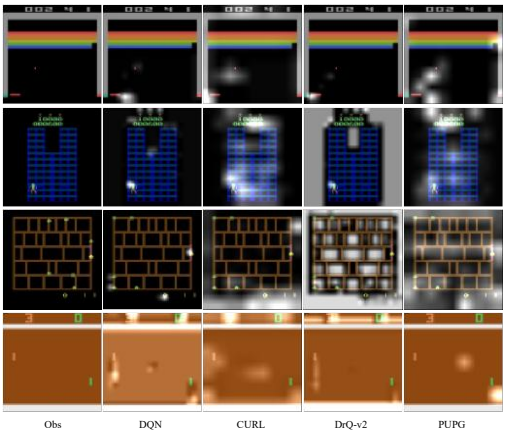

Fig. 7. Grad-CAM visualizations of compared VRL methods on 5 Atari games.

6.7 Trade-Off between Complexity and Performance

While higher order expansion theoretically improve performance, they increase complexity (1^{st}: $T(n^3)$, 2^{nd}: $T(2n^3)$, 3^{rd}: $T(3n^3)$ without proportional gains. As shwon in **Fig. 8,** our experiments reveal that the 2^{nd}-order expansion achieves near 3^{rd}-order performance (within 2%) at lower cost, making it our preferred choice for balancing complexity and performance.

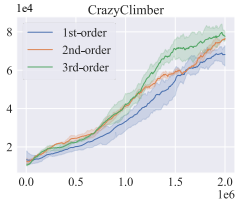

Fig. 8. Performance on CrazyClimber.

7 Conclusions

In this work, we propose PUPG, a novel positive-unlabeled policy-guided contrastive RL method for VRL, achieving SOTA sample efficiency across diverse benchmarks. Our work is the first to analyze VRL representations from an information-theoretic perspective. However, PUPG requires careful tuning of visual augmentation parameters to avoid suboptimal training. We hope our approach inspires further advances in VRL.

References

1. Sun, C., Wu, B., Wang, R., et al.: Intrinsic motivated multi-agent communication. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1668–1670. (2021)
2. Vinyals, O., Babuschkin, I., Czarnecki, W.M., et al.: Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature 575, 350–354. (2019)
3. Sun, C., Qiang, W., Li, J.: Demo:scdrl: Scalable and customized distributed reinforcement learning system. In: 44th IEEE International Conference on Distributed Computing Systems, ICDCS 2024, pp. 1404–1407. IEEE, Jersey City, NJ, USA (2024a)
4. Shao, K. et al.: A Survey of Deep Reinforcement Learning in Video Games. CoRR abs/1912.10944 (2019).
5. Li, Z. et al.: A Comprehensive Review of Multi-Agent Reinforcement Learning in Video Games. TechRxiv (2025).
6. Sun, C. et al.: M2I2: Learning Efficient Multi-Agent Communication via Masked State Modeling and Intention Inference. CoRR abs/2501.00312 (2025).
7. Sun, C. et al.: Revisiting Communication Efficiency in Multi-Agent Reinforcement Learning from the Dimensional Analysis Perspective. CoRR abs/2501.02888 (2025).
8. Sun, C., Zang, Z., Li, J., et al.: T2mac: Targeted and trusted multi-agent communication through selective engagement and evidence-driven integration. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38(13), 15154–15163. (2024)
9. Andrychowicz, O.M., Baker, B., Chociej, M., et al.: Learning dexterous in-hand manipulation. The International Journal of Robotics Research 39, 3–20. (2020)
10. Yarats, D., Fergus, R., Lazaric, A., Pinto, L.: Mastering visual continuous control: Improved data-augmented reinforcement learning. In: The Tenth International Conference on Learning Representations, ICLR 2022. OpenReview.net, Virtual Event (2022)
11. Laskin, M., Srinivas, A., Abbeel, P., et al.: CURL: contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, pp. 5639–5650. PMLR, Virtual Event (2020b)
12. Hessel, M., Modayil, J., van Hasselt, H., et al.: Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 3215–3222. AAAI Press, New Orleans, Louisiana, USA (2018)
13. Li, J., Zang, Z., Ji, Q., et al.: Rethinking generalizability and discriminability of self-supervised learning from evolutionary game theory perspective. (2024)
14. Li, J., Qiang, W., Zheng, C., et al.: Modeling multiple views via implicitly preserving global consistency and local complementarity. IEEE Trans. Knowl. Data Eng. 35, 7220–7238. (2023)
15. Chen, T. et al.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML 2020). Proceedings of Machine Learning Research, vol. 119, pp. 1597-1607. PMLR (2020).
16. He, K., Fan, H., Wu, Y., et al.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, pp. 9726–9735. Computer Vision Foundation / IEEE, Seattle, WA, USA (2020)
17. Grill, J., Strub, F., Altch´e, F., et al.: Bootstrap your own latent - A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems 33, NeurIPS 2020, pp. 1–15. NeurIPS Foundation, Virtual Event (2020)
18. Zbontar, J., Jing, L., Misra, I., et al.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, pp. 12310–12320. PMLR, Virtual Event (2021)

19. Li, J., Qiang, W., Zheng, C., et al.: Metaug: Contrastive learning via meta feature augmentation. In: Proceedings of the 39th International Conference on Machine Learning, ICML 2022, pp. 12964–12978. PMLR, Baltimore, Maryland, USA (2022c)
20. Yarats, D. et al.: Improving Sample Efficiency in Model-Free Reinforcement Learning from Images. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), pp. 10674-10681. AAAI Press (2021).
21. Schwarzer, M. et al.: Data-Efficient Reinforcement Learning with Self-Predictive Representations. In: 9th International Conference on Learning Representations (ICLR). OpenReview.net (2021).
22. Machado, M.C. et al.: Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. Journal of Artificial Intelligence Research 61, 523–562 (2018).
23. Tassa, Y., Doron, Y., Muldal, A., et al.: Deepmind control suite. CoRR abs/1801.00690. (2018)
24. Kim, M., Rho, K., Kim, Y., et al.: Action-driven contrastive representation for reinforcement learning. In: 17th International Proceedings on Plos One, pp. e0265456. Publisher, Location (2022)
25. Choi, H., Lee, H., Song, W., et al.: Local-Guided Global: Paired Similarity Representation for Visual Reinforcement Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, pp. 15072–15082. IEEE, Vancouver, BC, Canada (2023)
26. Liu, S., Zhang, X.S., Li, Y., et al.: On the Data-Efficiency with Contrastive Image Transformation in Reinforcement Learning. In: Proceedings of the 11th International Conference on Learning Representations, ICLR 2023. OpenReview.net, Kigali, Rwanda (2023)
27. Haarnoja, T. et al.: 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). Proceedings of Machine Learning Research, vol. 80, pp. 1856-1865. PMLR (2018).
28. Nguyen, T.X. et al.: Sample-efficient Reinforcement Learning Representation Learning with Curiosity Contrastive Forward Dynamics Model. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3471–3477. IEEE (2021).
29. Sun, C. et al.: CCLF: A Contrastive-Curiosity-Driven Learning Framework for Sample-Efficient Reinforcement Learning. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), pp. 3444-3450. ijcai.org (2022).
30. You, B. et al.: Integrating Contrastive Learning with Dynamic Models for Reinforcement Learning from Images. Neurocomputing 476, 102-114 (2022).
31. Yu, T. et al.: Mask-based Latent Reconstruction for Reinforcement Learning. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022). (2022).