

CURLing the Dream: Contrastive Representations for World Modeling in Reinforcement Learning

Victor A. Kich^{1*}, Jair A. Bottega¹, Raul Steinmetz², Ricardo B. Grando³, Ayano Yorozu¹ and Akihisa Ohya¹

¹Intelligent Robot Laboratory, University of Tsukuba,
 Tsukuba, Japan (victorkich98@gmail.com) * Corresponding author

²Technology Center, Federal University of Santa Maria,
 Santa Maria, Brazil (rsteinmetz@inf.ufsm.br)

³Robotics and AI Lab, Technological University of Uruguay,
 Rivera, Uruguay (ricardo.bedin@utec.edu.uy)

Abstract: In this work, we present Curled-Dreamer, a novel reinforcement learning algorithm that integrates contrastive learning into the DreamerV3 framework to enhance performance in visual reinforcement learning tasks. By incorporating the contrastive loss from the CURL algorithm and a reconstruction loss from autoencoder, Curled-Dreamer achieves significant improvements in various DeepMind Control Suite tasks. Our extensive experiments demonstrate that Curled-Dreamer consistently outperforms state-of-the-art algorithms, achieving higher mean and median scores across a diverse set of tasks. The results indicate that the proposed approach not only accelerates learning but also enhances the robustness of the learned policies. This work highlights the potential of combining different learning paradigms to achieve superior performance in reinforcement learning applications.

Keywords: Deep Reinforcement Learning, Contrastive Learning, Model-based, Representation Learning

1. INTRODUCTION

Reinforcement learning (RL) has achieved promising results in addressing complex tasks through deep neural networks [1], [2], [3]. However, the challenge of learning from high-dimensional visual inputs persists. Traditional approaches often struggle with sample efficiency and the quality of the learned representations. Recent advancements, such as the DreamerV3 algorithm, have shown promise by employing latent dynamics models to predict future states and rewards, thereby enhancing the efficiency of visual RL tasks [4]. Despite these improvements, there remains potential for enhancing representation learning and overall performance in RL algorithms [5], [1].

Contrastive learning has emerged as a powerful technique for unsupervised representation learning, demonstrating impressive results across various domains such as computer vision and natural language processing. One notable algorithm, CURL, applies contrastive learning principles to RL by maximizing the agreement between different augmentations of the same state while minimizing the agreement between different states [6], [7], [8], [9], [10]. This approach has proven effective in enhancing the performance and robustness of RL algorithms [11], [12].

We propose an enhancement to the DreamerV3 algorithm by integrating the contrastive loss from CURL. Our method, termed Curled-Dreamer, aims to improve the encoder's ability to capture informative and discriminative features from visual inputs. The **Curled-Dreamer**¹ algorithm modifies the training procedure of the encoder in DreamerV3 to include a contrastive loss that promotes the similarity of representations for positive pairs (different augmentations of the same state) and dissimilarity for neg-

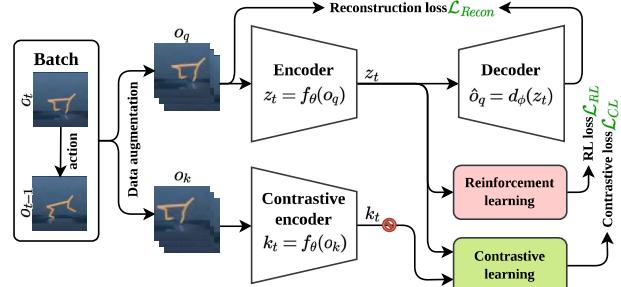


Fig. 1.: Overview of the Curled-Dreamer algorithm, illustrating the integration of contrastive and reconstruction losses to enhance visual representation learning.

ative pairs (augmentations of different states) [4], [6], [7], [8], [9]. This combined approach leverages the predictive power of DreamerV3 while improving the quality of state representations through contrastive learning. The system diagram can be seen in Fig. 1.

The contributions of this work are twofold:

- We introduced the Curled-Dreamer algorithm, which integrates contrastive learning into the DreamerV3 framework to enhance visual reinforcement learning.
- Through extensive experiments, we demonstrate that Curled-Dreamer outperforms several state-of-the-art algorithms on a suite of DeepMind Control Suite tasks, achieving higher mean and median scores [13], [4], [6], [5], [1], [7], [8], [9].

These contributions highlight the potential of combining different learning paradigms to achieve superior performance in reinforcement learning applications.

2. RELATED WORKS

Contrastive learning has emerged as a transformative approach in RL by enhancing the quality of learned repre-

¹Code available at: <https://github.com/victorkich/curled-dreamer>

sentations and improving sample efficiency. Eysenbach et al. (2022) demonstrated that contrastive representation learning methods can be directly applied as RL algorithms, achieving higher success rates in goal-conditioned tasks and outperforming prior methods on image-based tasks without additional data augmentation or auxiliary objectives [14]. Yang et al. (2023) proposed a novel unsupervised skill discovery method through contrastive learning, which maximizes mutual information between behaviors based on the same skill, leading to the generation of diverse and far-reaching skills [15].

In the realm of model-based RL, Ma et al. (2020) introduced Contrastive Variational Reinforcement Learning (CVRL), which leverages contrastive learning to maximize mutual information between latent states and observations. This approach has shown robustness in tasks with complex visual observations, outperforming state-of-the-art generative world models [16]. Kim et al. (2022) developed the Action-Driven Auxiliary Task (ADAT) method, which focuses on essential features for decision-making and significantly enhances feature learning and performance in standard benchmarks [17].

Yuan and Lu (2022) addressed the challenge of distribution mismatch in offline meta-reinforcement learning by proposing a contrastive learning framework that improves task representation robustness and generalization to out-of-distribution behavior policies [18]. Shen et al. (2021) introduced Sequential and Dynamic Constraint Contrastive Reinforcement Learning (SDCRL), incorporating sequential information and dynamic transitions to improve sample efficiency and performance on complex tasks [19].

Li et al. (2023) introduced Model-Enhanced Contrastive Reinforcement Learning (MCRL) for sequential recommendation, addressing data sparsity and overestimation issues, and significantly outperforming existing methods [20]. Qiu et al. (2022) proposed Contrastive UCB, an algorithm integrating contrastive learning with Upper Confidence Bound (UCB) methods for improved sample efficiency and representation learning in online RL settings [21].

Lastly, Wang and Hu (2023) developed lightweight and effective DRL algorithms incorporating contrastive learning to balance new and old experience trajectories, achieving state-of-the-art performance in PyBullet robotics environments [22]. These studies collectively highlight the potential of contrastive learning to advance RL by enhancing representation quality, sample efficiency, and generalization capabilities.

3. METHODOLOGY

In this work, we propose an enhancement to the DreamerV3 algorithm by integrating the contrastive loss from the CURL algorithm, aiming to improve the encoder's ability to capture informative and discriminative features from visual inputs. Curled-Dreamer, modifies the training procedure of the encoder in DreamerV3 to include a

contrastive loss that promotes the similarity of representations for positive pairs (different augmentations of the same state) and dissimilarity for negative pairs (augmentations of different states). Additionally, we incorporate a reconstruction loss in the output of the decoder to further enhance the quality of the learned representations.

The Curled-Dreamer algorithm operates through the following steps. First, we perform data augmentation on each state observation by performing random crops. Let $\mathcal{S} = \{s_i\}_{i=1}^N$ be a set of original states, and $\mathcal{A} = \{a_i\}_{i=1}^N$ be the corresponding augmented states. This step aims to improve the robustness of the representations learned by the encoder. Next, the encoder f_θ is trained using both the objectives of DreamerV3 and a contrastive loss. The contrastive loss is computed using the InfoNCE (Information Noise Contrastive Estimation) loss [23], defined as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{\mathcal{A}} \left[\log \frac{\exp(\text{sim}(f_\theta(a_i), f_\theta(a_i^+))/\tau)}{\sum_{a_j \in \mathcal{A}} \exp(\text{sim}(f_\theta(a_i), f_\theta(a_j))/\tau)} \right], \quad (1)$$

where a_i^+ is the positive pair of a_i (i.e., another augmentation of the same state), $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, and τ is a temperature parameter. This contrastive loss encourages the encoder to produce similar representations for different augmentations of the same state and dissimilar representations for different states.

The representations $z_t = f_\theta(s_t)$ obtained from the encoder are then used as inputs to the latent dynamics model. The latent dynamics model g_ϕ and reward predictor r_ψ are trained to predict future states and rewards, respectively. The loss function for training the latent dynamics model and the reward predictor is given by:

$$\mathcal{L}_{\text{dynamics}} = \mathbb{E} \left[\sum_{t=1}^T (\|g_\phi(z_t, a_t) - z_{t+1}\|^2 + \|r_\psi(z_t, a_t) - r_t\|^2) \right], \quad (2)$$

where z_t is the latent state representation at time t , a_t is the action taken at time t , and r_t is the reward received at time t . This loss ensures that the latent dynamics model accurately predicts the future latent states and rewards based on the current latent state and action.

The encoder's output is also fed into a decoder d_η to reconstruct the original input, promoting the learning of more detailed and accurate representations. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{reconstruction}} = \mathbb{E} [\|d_\eta(f_\theta(s_t)) - s_t\|^2]. \quad (3)$$

The policy π_θ is optimized using the predictions from the latent dynamics model. The policy optimization follows the actor-critic method used in DreamerV3. The policy loss is given by:

$$\mathcal{L}_{\text{policy}} = -\mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^T \gamma^t r_\psi(z_t, a_t) \right], \quad (4)$$

Algorithm 1 Curled-Dreamer Algorithm

```
1: Input: Set of states  $\mathcal{S}$ , actions  $\mathcal{A}$ , hyperparameters  $\lambda_1, \lambda_2, \lambda_3, \tau$ 
2: Initialize: Encoder  $f_\theta$ , latent dynamics model  $g_\phi$ , reward predictor  $r_\psi$ , decoder  $d_\eta$ , policy  $\pi_\theta$ 
3: for each iteration do
4:   Sample a batch of states  $\{s_i\}_{i=1}^N$  from  $\mathcal{S}$ 
5:   Generate augmented states  $\{a_i\}_{i=1}^N$  from  $\mathcal{A}$ 
6:   Compute latent representations  $\{z_i = f_\theta(a_i)\}_{i=1}^N$ 
7:   Compute InfoNCE loss  $\mathcal{L}_{\text{InfoNCE}}$  using Eq. (1)
8:   Compute dynamics loss  $\mathcal{L}_{\text{dynamics}}$  using Eq. (2)
9:   Compute reconstruction loss  $\mathcal{L}_{\text{reconstruction}}$  using Eq. (4)
10:  Compute policy loss  $\mathcal{L}_{\text{policy}}$  using Eq. (3)
11:  Compute total loss  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{policy}} + \lambda_1 \mathcal{L}_{\text{dynamics}} + \lambda_2 \mathcal{L}_{\text{InfoNCE}} + \lambda_3 \mathcal{L}_{\text{reconstruction}}$ 
12:  Update encoder  $f_\theta$ , latent dynamics model  $g_\phi$ , reward predictor  $r_\psi$ , decoder  $d_\eta$ , and policy  $\pi_\theta$  using gradient
    descent
13: end for
14: Output: Trained encoder  $f_\theta$ , latent dynamics model  $g_\phi$ , reward predictor  $r_\psi$ , decoder  $d_\eta$ , and policy  $\pi_\theta$ 
```

where γ is the discount factor. This loss function maximizes the expected cumulative reward over time, guiding the policy to select actions that lead to higher rewards.

The overall training objective for Curled-Dreamer is a combination of the contrastive loss, the dynamics model loss, the policy loss, and the reconstruction loss. The total loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{policy}} + \lambda_1 \mathcal{L}_{\text{dynamics}} + \lambda_2 \mathcal{L}_{\text{InfoNCE}} + \lambda_3 \mathcal{L}_{\text{reconstruction}}, \quad (5)$$

where λ_1, λ_2 , and λ_3 are hyperparameters that balance the contributions of the dynamics, contrastive, and reconstruction losses, respectively. By optimizing this combined loss, Curled-Dreamer leverages the predictive power of DreamerV3 while enhancing the quality of state representations through contrastive and reconstruction learning, leading to more robust and efficient learning.

The pseudocode for the Curled-Dreamer algorithm is provided in Algorithm 1.

To summarize, Curled-Dreamer integrates contrastive learning and reconstruction loss into the DreamerV3 framework to improve visual representation learning in reinforcement learning tasks. By combining data augmentation, contrastive loss, reconstruction loss, and traditional reinforcement learning objectives, the proposed method enhances the encoder’s ability to capture discriminative features, thereby improving the overall performance and robustness of the learned policies.

4. EXPERIMENTS

In this section, we present the experimental setup and the results obtained from evaluating the proposed Curled-Dreamer algorithm. We describe the environments used, the hyperparameters adopted, and provide an in-depth analysis of the results.

We evaluated the performance of Curled-Dreamer on a suite of 20 tasks from the DeepMind Control Suite (DMC) [13]. The DMC provides a diverse set of continuous control tasks designed to test various aspects of reinforcement

learning algorithms, including balance, locomotion, and manipulation. The selected tasks offers a comprehensive evaluation of the algorithm’s capability to handle various control challenges, including high-dimensional observations and continuous action spaces.

To ensure a fair comparison, we employed the same hyperparameters as used in the original DreamerV3 paper [4]. The key hyperparameters are as follows: the learning rate for the encoder, dynamics model, reward predictor, decoder, and policy is set to 3×10^{-4} ; a batch size of 50 is used for training; the discount factor (γ) is set to 0.99; the sequence length for training the recurrent models is set to 50; the temperature parameter (τ) for the InfoNCE loss is set to 0.1; the weights for the contrastive loss (λ_2), reconstruction loss (λ_3), and dynamics loss (λ_1) are all set to 1.0. These hyperparameters were chosen based on extensive tuning performed in the DreamerV3 paper and have been shown to provide robust performance across a wide range of tasks.

We evaluated the performance of Curled-Dreamer using the average return over 1 million environment steps. This metric provides a comprehensive assessment of the learning efficiency and policy performance. For each task, we report the mean and median scores achieved by our algorithm, and we compare these results against several state-of-the-art baseline algorithms, including PPO [24], SAC [25], CURL [6], DrQ-v2 [26], and previous versions of DreamerV3 [4].

5. RESULTS

Overall, Curled-Dreamer demonstrated substantial improvements across a variety of tasks, showcasing its versatility and robustness. The algorithm achieved higher mean and median scores across the suite of tasks, indicating a significant enhancement in learning efficiency and policy performance. Some environments are shown in Fig. 2. While the results are summarized in Table 1 and depicted in Fig. 3.

Particularly outstanding results were observed in tasks that involve complex dynamical systems and high-

Table 1.: DMC scores for visual inputs at 1M frames.

Task	PPO[24]	SAC[25]	CURL[6]	DrQ-v2[26]	DV3 (2023)[4]	DV3 (2024)[4]	Curled-Dreamer
Train Steps	1M	1M	1M	1M	1M	1M	1M
Acrobot Swingup	3	4	4	166	210	229	328
Ball In Cup Catch	829	176	970	928	957	972	960
Cartpole Balance	516	937	980	992	996	993	996
Cartpole Balance Sparse	881	956	999	987	1000	964	996
Cartpole Swingup	290	706	771	863	819	861	859
Cartpole Swingup Sparse	1	149	373	773	792	759	769
Cheetah Run	95	20	502	716	728	836	867
Finger Spin	118	291	880	862	818	589	617
Finger Turn Easy	253	200	340	525	787	878	767
Finger Turn Hard	79	94	231	247	810	904	896
Hopper Hop	0	0	164	221	369	227	350
Hopper Stand	4	5	777	903	900	903	871
Pendulum Swingup	1	592	413	843	806	744	797
Quadruped Run	88	54	149	450	352	617	807
Quadruped Walk	112	49	121	726	352	811	894
Reacher Easy	487	67	689	944	898	951	934
Reacher Hard	94	7	472	670	499	862	741
Walker Run	30	27	360	539	757	684	719
Walker Stand	161	143	486	978	976	976	975
Walker Walk	87	40	822	768	955	961	956
Task Mean	94	81	479	770	808	861	863
Task Median	206	226	525	705	739	786	805

dimensional observations. For instance, Curled-Dreamer excelled in the Acrobot Swingup task, achieving a score of 328, which is nearly double the score of the next best baseline. This improvement can be attributed to the enhanced representation learning facilitated by the contrastive and reconstruction losses, enabling the model to better capture the intricate system dynamics.

In manipulation tasks such as Finger Turn Easy and Finger Turn Hard, Curled-Dreamer achieved scores of 767 and 896, respectively. These results surpass those of the previous best-performing models, indicating that the enhanced representation learning contributes to better manipulation and control capabilities. The ability to capture fine-grained details from visual inputs likely played a crucial role in these tasks, where precise movements are required.

Locomotion tasks, including Cheetah Run and Quadruped Walk, also benefited significantly from the proposed enhancements. In the Cheetah Run task, Curled-Dreamer achieved a score of 867, outperforming the previous best score of 836. The improved representation learning allowed the model to better understand and predict the agent’s movements, resulting in more effective and coordinated locomotion. Similarly, in the Quadruped Walk task, Curled-Dreamer achieved an impressive score

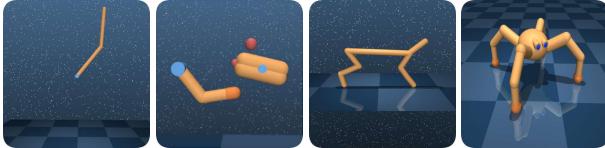


Fig. 2.: Rendered demonstration for the following tasks: Acrobot Swingup, Finger Turn, Cheetah, and Quadruped.

of 894, showcasing its ability to generalize across different types of locomotion challenges.

Moreover, tasks with sparse rewards, such as Cartpole Balance Sparse and Cartpole Swingup Sparse, demonstrated the robustness of Curled-Dreamer in learning stable policies despite limited feedback. The algorithm achieved near-perfect scores in these tasks, indicating its proficiency in extracting valuable information from high-dimensional observations and learning effective control strategies even in challenging environments.

The results also highlight the effectiveness of the combined learning paradigms in enhancing the model’s performance. The contrastive loss improved the encoder’s ability to capture informative and discriminative features from visual inputs, while the reconstruction loss promoted the learning of more detailed and accurate representations. These components, along with the robust training procedure and well-tuned hyperparameters, contributed to the overall success of Curled-Dreamer.

In conclusion, the experimental results provide strong evidence of the effectiveness of Curled-Dreamer in addressing various reinforcement learning tasks. The combination of DreamerV3’s predictive capabilities and the enhanced representation learning from CURL makes Curled-Dreamer a new option for model-based reinforcement learning.

6. CONCLUSION

In this paper, we introduced Curled-Dreamer, an enhanced reinforcement learning algorithm that integrates contrastive learning into the DreamerV3 framework. By

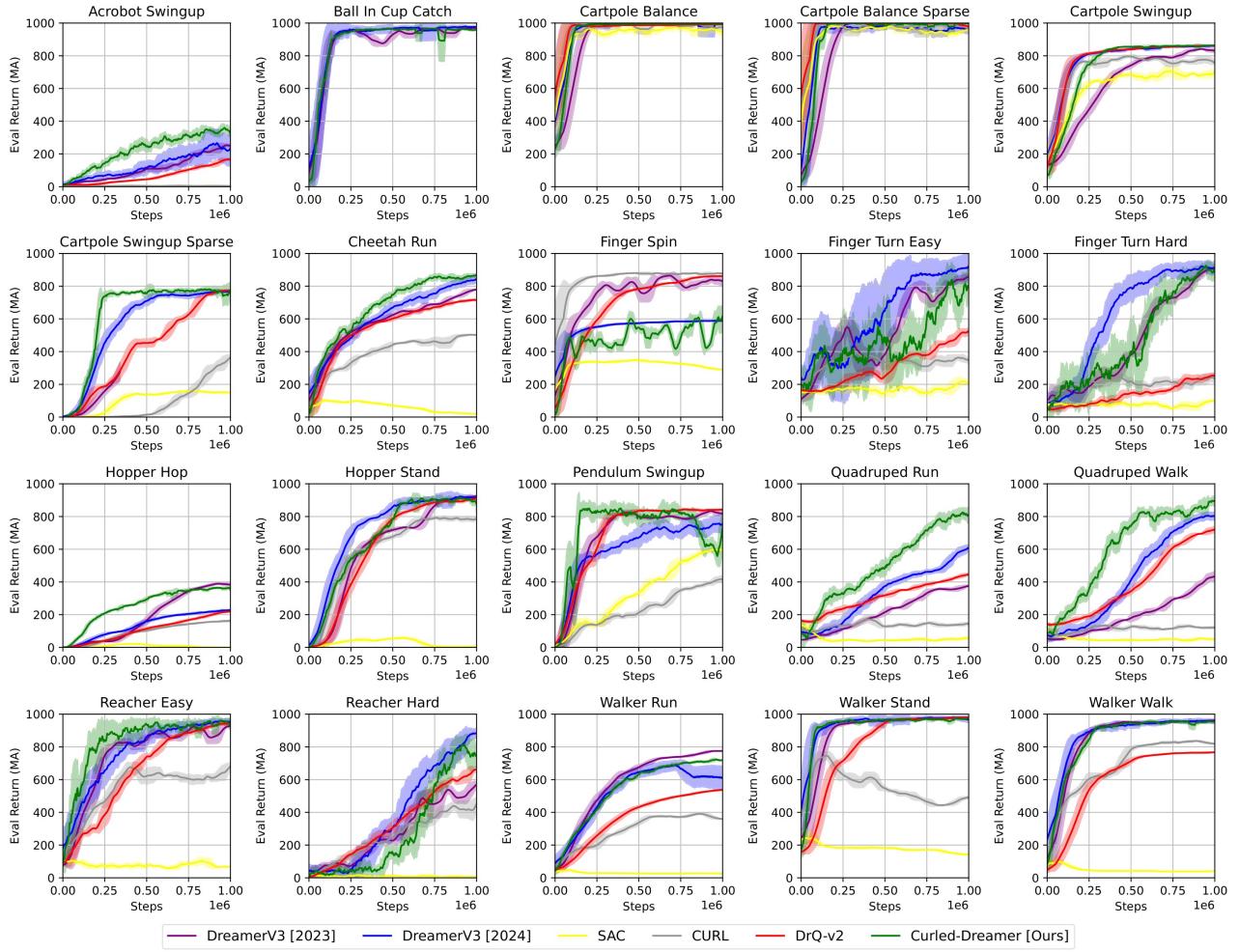


Fig. 3.: Evaluation returns on various DMC tasks for 1M steps.

incorporating the contrastive loss from CURL and a reconstruction loss, Curled-Dreamer improves the quality of learned representations, resulting in enhanced performance and robustness in visual reinforcement learning tasks. Our experiments on the DeepMind Control Suite demonstrated that Curled-Dreamer consistently outperforms or matches the performance of state-of-the-art algorithms, highlighting its effectiveness across a diverse set of tasks.

The results indicate that the representation learning achieved through contrastive and reconstruction losses is key to Curled-Dreamer’s performance. These components enable the model to capture more informative and discriminative features from high-dimensional visual inputs, leading to better policy performance. The robust training procedure and well-tuned hyperparameters further contribute to the algorithm’s success, ensuring stability and efficiency in learning.

Furthermore, the addition of parameters, such as the use of two encoders and the matrix W , likely contributed to the improved results by allowing the model to capture more complex features and relationships within the data. This increase in model capacity, enhanced Curled-Dreamer’s ability to learn and generalize from visual in-

puts.

Future work will explore parallelism and remove the decoder from the model to make the algorithm more robust. Additionally, implementing Curled-Dreamer using JAX instead of PyTorch could make it faster. Although there is no performance comparison in this paper, our approach, even with more extensive learning, is currently slower than the compared state-of-the-art methods.

Overall, Curled-Dreamer showcases the potential of combining different learning paradigms to achieve superior results in reinforcement learning. Investigating the scalability and generalization capabilities of Curled-Dreamer in more complex environments could provide valuable insights into its applicability in real-world scenarios.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship. We are also grateful for the support provided by the Human-Centered AI program at the University of Tsukuba.

REFERENCES

- [1] V. Mnih and et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [2] R. B. Grando, R. Steinmetz, V. A. Kich, A. H. Kolling, P. M. Furik, J. C. de Jesus, B. V. Guterres, D. T. Gamarra, R. S. Guerra, and P. L. Drews-Jr, “Improving generalization in aerial and terrestrial mobile robots control through delayed policy learning,” *arXiv preprint arXiv:2406.01952*, 2024.
- [3] V. A. Kich, J. C. de Jesus, R. B. Grando, A. H. Kolling, G. V. Heisler, and R. d. S. Guerra, “Deep reinforcement learning using a low-dimensional observation filter for visual complex video game playing,” *arXiv preprint arXiv:2204.11370*, 2022.
- [4] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, “Mastering diverse domains through world models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [5] T. P. Lillicrap and et al., “Continuous control with deep reinforcement learning,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [6] M. Laskin, A. Srinivas, and P. Abbeel, “Curl: Contrastive unsupervised representations for reinforcement learning,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- [7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- [8] K. He and et al., “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [9] J.-B. Grill and et al., “Bootstrap your own latent: A new approach to self-supervised learning,” in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020.
- [10] J. C. de Jesus, V. A. Kich, A. H. Kolling, R. B. Grando, R. S. Guerra, and P. L. Drews, “Depthcuprl: Depth-imaged contrastive unsupervised prioritized representations in reinforcement learning for mapless navigation of unmanned aerial vehicles,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 579–10 586.
- [11] A. Stooke, L. Pinto, and et al., “Rlbench: The robot learning benchmark,” in *Proceedings of the Conference on Robot Learning (CoRL)*, 2021.
- [12] M. Schwarzer and et al., “Data-efficient reinforcement learning with self-predictive representations,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [13] Y. Tassa, Y. Doron, and et al., “Deepmind control suite,” *arXiv preprint arXiv:1801.00690*, 2018.
- [14] B. Eysenbach, T. Zhang, S. Levine, and R. R. Salakhutdinov, “Contrastive learning as goal-conditioned reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 603–35 620, 2022.
- [15] R. Yang, C. Bai, H. Guo, S. Li, B. Zhao, Z. Wang, P. Liu, and X. Li, “Behavior contrastive learning for unsupervised skill discovery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [16] X. Ma, S. Chen, D. Hsu, and W. Lee, “Contrastive variational model-based reinforcement learning for complex observations,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- [17] M. Kim, K. Rho, Y.-D. Kim, and K. Jung, “Action-driven contrastive representation for reinforcement learning,” *PLoS ONE*, vol. 17, no. 2, p. e0263618, 2022.
- [18] H. Yuan and Z. Lu, “Robust task representations for offline meta-reinforcement learning via contrastive learning,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- [19] W. Shen, L. Yuan, J. Huang, S. Gao, Y. Huang, and Y. Yu, “Sequential and dynamic constraint contrastive learning for reinforcement learning,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.
- [20] C. Li, Z. Yang, J. Zhang, J. Wu, D. Wang, X. He, and X. Wang, “Model-enhanced contrastive reinforcement learning for sequential recommendation,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [21] S. Qiu, L. Wang, C. Bai, Z. Yang, and Z. Wang, “Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning,” in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022.
- [22] D. Wang and M. Hu, “Contrastive learning methods for deep reinforcement learning,” *IEEE Access*, vol. 11, pp. 97 107–97 117, 2023.
- [23] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” in *32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.
- [24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” in *International Conference on Machine Learning (ICML)*, 2017.
- [25] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International Conference on Machine Learning (ICML)*, 2018.
- [26] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, “Mastering visual continuous control: Improved data-augmented reinforcement learning,” in *International Conference on Learning Representations (ICLR)*, 2021.