

Efficient Reinforcement Learning Through Adaptively Pretrained Visual Encoder

Yuhan Zhang^{12*}, Guoqing Ma^{13*}, Guangfu Hao¹², Liangxuan Guo¹³, Yang Chen¹⁴, Shan Yu^{134†}

¹Laboratory of Brain Atlas and Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

³School of Future Technology, University of Chinese Academy of Sciences

⁴Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Chinese Academy of Sciences
{zhangyuhan2022, maguoqing2022, haoguangfu2021, guoliangxuan2021, yang.chen, shan.yu}@ia.ac.cn

Abstract

While Reinforcement Learning (RL) agents can successfully learn to handle complex tasks, effectively generalizing acquired skills to unfamiliar settings remains a challenge. One of the reasons behind this is the visual encoders used are task-dependent, preventing effective feature extraction in different settings. To address this issue, recent studies have tried to pretrain encoders with diverse visual inputs in order to improve their performance. However, they rely on existing pretrained encoders without further exploring the impact of pretraining period. In this work, we propose APE: efficient reinforcement learning through Adaptively Pretrained visual Encoder—a framework that utilizes adaptive augmentation strategy during the pretraining phase and extracts generalizable features with only a few interactions within the task environments in the policy learning period. Experiments are conducted across various domains, including DeepMind Control Suite, Atari Games and Memory Maze benchmarks, to verify the effectiveness of our method. Results show that mainstream RL methods, such as DreamerV3 and DrQ-v2, achieve state-of-the-art performance when equipped with APE. In addition, APE significantly improves the sampling efficiency using only visual inputs during learning, approaching the efficiency of state-based method in several control tasks. These findings demonstrate the potential of adaptive pretraining of encoder in enhancing the generalization ability and efficiency of visual RL algorithms.

Introduction

Deep Reinforcement Learning (Deep RL) has made great advances in recent years. Notable algorithms such as MuZero (Schrittwieser et al. 2019), Player of Games (Schmid et al. 2021) and ReBeL (Brown et al. 2020) have been proposed to solve many challenging decision making problems. While these advances have primarily focused on state-based inputs, significant progress has also been made in visual RL, i.e., leveraging image inputs for policy learning (Srinivas, Laskin, and Abbeel 2020; Hafner et al. 2019, 2020, 2023; Kostrikov, Yarats, and Fergus 2020).

However, visual RL agents learning from these high-dimensional observations suffer from problems of low effi-

*These authors contributed equally.

†Corresponding author.

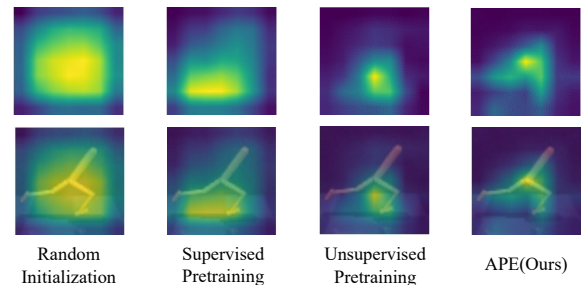


Figure 1: Visualization of ResNet-18 model with different pretraining strategy using LayerCAM (Jiang et al. 2021), which indicates that APE is able to extract more precise outline of the Walker than other initialization settings. The first row displays the pure feature maps, which are also presented together with the image in the second row.

ciency and often overfitting to specific environments (Song et al. 2019; Srinivas, Laskin, and Abbeel 2020). Since the performance of these agents depends heavily on the quality of extracted features, the critical role of enhancing visual encoders has been highlighted in both model-free and model-based algorithms (Yarats et al. 2019; Hafner et al. 2019; Poudel et al. 2023).

In visual RL, various approaches have been explored to improve representation learning, among which data augmentations are often used to increase data diversity (Wang et al. 2020; Raileanu et al. 2021; Liu et al. 2022, 2023). The challenge lies in extracting generalizable features rather than focusing on task-specific details, leading to difficulties in transferring learned skills to unseen scenarios (Lee et al. 2019; Laskin et al. 2020).

One promising direction is to exploit cross-domain knowledge learned by pretrained models (Shah and Kumar 2021; Yuan et al. 2022), which has shown great success in improving data efficiency and generalization ability in recent deep learning (Devlin et al. 2019; Baevski et al. 2020). In computer vision, since these models have typically been trained on extensive sets of natural images, their features inherently possess general knowledge about the world (Hu et al. 2023). This approach has the potential to enable RL agents to extract useful features more effectively, enhancing their ability to learn and generalize across different domains.

Unsupervised learning, e.g., contrastive learning, is particularly advantageous in this regard as it enables pretrained models to extract meaningful features from unlabeled visual data, effectively addressing the issue of data scarcity and high labeling costs (He et al. 2020; Chen et al. 2020a).

Nevertheless, current RL methods simply implement existing pretrained models as visual encoders and augment observations in the downstream policy learning period (Shah and Kumar 2021; Hu et al. 2023). As illustrated in Fig. 1, the features learned by image classification models with the prevailing pretraining strategies (shown in the left three columns) exhibit limited generalization capabilities. This also results in a lack of exploration of pretraining augmentations, which prove to be an important factor when applying pretrained encoders under great distribution shifts (Geirhos et al. 2021; Burns et al. 2023).

Given this, here we propose APE, a framework where the RL agent learns efficiently through **Adaptively Pretrained visual Encoder**. This novel framework uses an adaptive closed-loop augmentation strategy in contrastive pretraining to learn transferable representations from a wide range of real-world images. Comparison in Fig. 1 indicates that APE helps to extract more generalizable features than other pretraining strategies. In addition, it works efficiently, requiring minimal interactions with the targeted environment during policy learning period. We evaluate our method on various challenging visual RL domains, including DeepMind Control (DMC) Suite (Tassa et al. 2018), the Atari 100K benchmark (Bellemare et al. 2012), and Memory Maze (Paukonis, Lillicrap, and Hafner 2022). Experiments demonstrate that APE significantly improves the sampling efficiency and performance of the base RL method. Interestingly, we found that the real RL environments are not necessary to test the pretrained encoder. Linear probes, a common protocol for evaluating the quality of learned representations (Chen et al. 2020a), can serve as a useful metric to assess the quality of pretrained encoders quite effectively. The main contribution of this paper can be summarized as follows:

- We propose a cross-domain RL framework with a fixed encoder pretrained on a wide variety of natural images using adaptive augmentation adjustment. This helps to produce more generalizable representations for the downstream RL tasks.
- We demonstrate the generality of APE to both model-based and model-free methods, underscoring its adaptability and effectiveness in enhancing learning performance across diverse RL approaches.
- APE is developed without any auxiliary tasks or other sensory information during policy learning period, effectively decoupling the pretraining phase from subsequent behavior learning tasks. This simple yet powerful design contributes to APE’s superior performance on various visual RL benchmarks, approaching the performance of state-based Soft-Actor-Critic (SAC) (Harnoja et al. 2018) in several control tasks.

Related Works

Contrastive Learning

In computer vision (CV), contrastive learning has gained popularity for its ability to learn generalizable representations leveraging unlabeled images and videos (van den Oord, Li, and Vinyals 2018; Chen et al. 2020a; He et al. 2020). Prior studies have emphasized the pivotal role of data augmentation in facilitating unsupervised training (Asano, Rupprecht, and Vedaldi 2019; Gidaris, Singh, and Komodakis 2018; Henaff 2020). Experiments conducted in SimCLR approach (Chen et al. 2020a) highlight the significant impact of data augmentations, which is re-confirmed by MoCo (He et al. 2020) and its modification MoCo v2 (Chen et al. 2020b). AdDA (Zhang, Zhu, and Yu 2023) focuses on exploring the effect of dynamic adjustment on augmentation compositions, which enables the network to acquire more generalizable features. We adopt the feedback structure (Zhang, Zhu, and Yu 2023) in the pretraining period and implement it on a different network architecture, which proves to be more suitable for RL tasks (Yuan et al. 2022).

Representation Learning in RL

There are extensive works in RL studying the impact of representation learning (Gelada et al. 2019; Lin et al. 2020a; Liu et al. 2023), among which contrastive learning is often applied to acquire useful features (Zhan et al. 2020; Du, Gan, and Isola 2021; Schwarzer et al. 2021). CURL (Srinivas, Laskin, and Abbeel 2020) trains a visual representation encoder using contrastive loss, significantly improving sampling efficiency over prior pixel-based methods. Proto-RL (Yarats et al. 2021b) learns contrastive visual representations in dynamic RL environments without access to task-specific rewards. To make full use of context information, MLR (Yu et al. 2022) introduces mask-based reconstruction to promote contrastive representation learning in RL. However, prior methods rely completely on data collected in target environments, which limits their generalization to unseen scenarios and hinders their adaptability to new tasks or environments. It also leads to additional sampling costs. APE, on the other hand, is pretrained on a distribution of real-world samples that wider than what policy can provide.

Besides, the interpretability of extracted features is a key focus (Lin et al. 2020b; Delfosse et al. 2022, 2024), leading to improved performance and robustness of the agent. The efficiency gains of our method also result from a more interpretable encoder, aiding the agent in capturing key factors of observations in policy-making period.

Generalization for Image-Based RL

Since image augmentation has been successfully applied in CV for improving performance on object classification tasks, different approaches of transformation were investigated and incorporated in RL pipelines (Laskin et al. 2020; Kostrikov, Yarats, and Fergus 2020; Stooke et al. 2020). DrAC (Raileanu et al. 2021) contributes to the proper use of data augmentation for actor-critic algorithms and proposes an automatically selecting approach. SVEA (Hansen, Su, and Wang 2021) investigates the factors contributing to

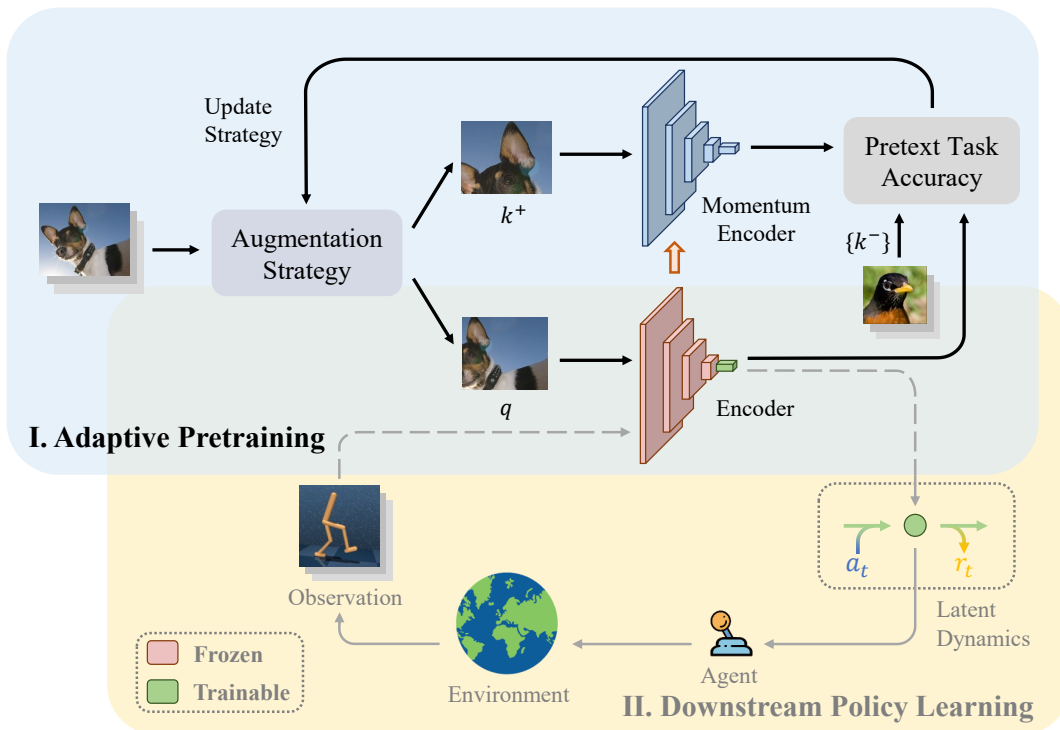


Figure 2: APE pipeline for MBRL. The training phase is divided into two parts, namely the Adaptive Pretraining period (within the blue area) and the Downstream Policy Learning period (within the yellow area). A wide variety of real-world images are augmented using an adaptive data augmentation strategy in the first period, which dynamically updates the sampling probability of each augmentation composition in the next pretraining epoch. In the second stage, the pretrained vision encoder is implemented in a generic RL framework as a perception module for the policy.

instability when employing augmentation within off-policy RL methods. DrQ (Kostrikov, Yarats, and Fergus 2020) together with DrQ-v2 (Yarats et al. 2021a) introduces a simple augmentation method for model-free RL algorithms utilizing input perturbations and regularization techniques, which we use to evaluate the generality of APE. However, most previous methods attach more importance to the policy training period and straightforwardly augment the observations of the target environments (Zhao et al. 2024). Thus, they fall short in providing the requisite data diversity, which is essential for generalization over large domain gaps (Yuan et al. 2022). On the contrary, APE leverages an adaptively pretrained encoder without neglecting the potential benefits of pretraining augmentation strategy in RL, which has been confirmed in recent studies for its effectiveness in enhancing RL performance (Burns et al. 2023).

Pretrained Visual Encoders for RL

Instead of training with expensive collected data, researches have also been made to bridge the domain gap between cross-domain datasets and the inputs of the target environments (Ma et al. 2022; Hu et al. 2023). Using a pretrained ResNet encoder, RRL (Shah and Kumar 2021) brings a straightforward approach to fuse extracted features into a standard RL pipeline. PIE-G (Yuan et al. 2022) further demonstrates the effectiveness of supervised pretrained

encoders by using early layer features from frozen models, with strongly augmented representations. By combining pretrained visual encoder and proprioceptive information, MVP outperforms supervised encoders in motor control tasks (Xiao et al. 2022). While pretrained models in aid of model-free RL have been studied, there lacks exploration on Model-Based Reinforcement Learning (MBRL) algorithms. These methods rely completely on reconstructed latents, thus further highlights the significance of representation learning (Poudel et al. 2023). Besides, extra tasks or sensory data are often needed during policy learning period while APE works without such intensive task-specific data.

Preliminaries

The proposed APE expands on both model-based and model-free RL methods. Detailed analyses are conducted on a mainstream MBRL framework, DreamerV3 (Hafner et al. 2023), which only learns from the representations extracted from original image observations. This integration allows APE to inherit DreamerV3’s generality, operating with fixed hyperparameters across various domains. This section provides an overall description of our MBRL Backbone.

Latent dynamics. The latent dynamics of DreamerV3 are modeled as a recurrent state space model (RSSM) which consists of the following five components:

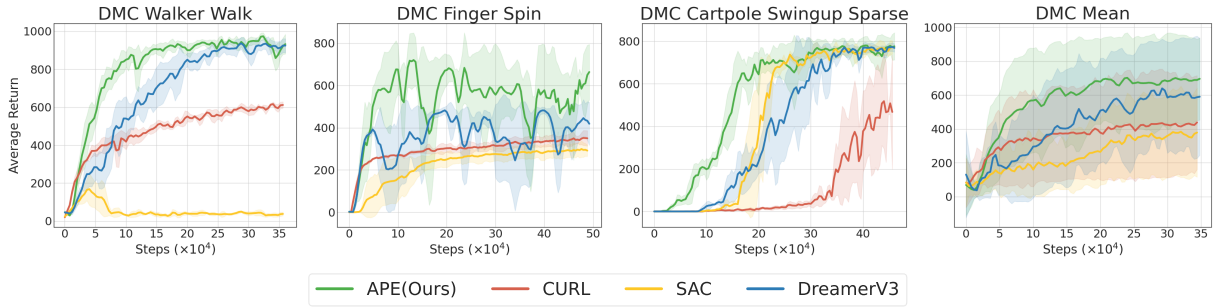


Figure 3: Training curves for DMC vision benchmarks.

$$\begin{aligned}
 \text{Encoder:} & \quad z_t \sim f_\theta(z_t | z_{t-1}, a_{t-1}, o_t) \\
 \text{Dynamics model:} & \quad \hat{z}_t \sim p_\theta^D(\hat{z}_t | z_{t-1}, a_{t-1}) \\
 \text{Reward predictor:} & \quad \hat{r}_t \sim p_\theta^R(\hat{r}_t | z_t, z_{t-1}, a_{t-1}) \\
 \text{Continue predictor:} & \quad \hat{c}_t \sim p_\theta^C(\hat{c}_t | z_t, z_{t-1}, a_{t-1}) \\
 \text{Decoder:} & \quad \hat{x}_t \sim g_\theta(\hat{x}_t | z_t, z_{t-1}, a_{t-1})
 \end{aligned} \quad (1)$$

Here the dynamics model is designed to predict the next latent representation \hat{z}_t , while the feature z_t generated by the encoder is used in the reward and continue predictor. The decoder using a convolutional neural network (CNN) helps in reconstructing visual inputs.

Agent learning. The actor-critic algorithm is employed to learn behaviors from the feature sequences predicted by the world model (Ha and Schmidhuber 2018). The actor aims to maximize the expected return R_t for each state s_t while the critic is trained to predict the return of each state s_t with the current action a_t . Given γ as the discount factor for the future rewards, the agent model are defined as follows:

$$\begin{aligned}
 \text{Actor:} \quad a_t & \sim \pi_\phi(a_t | s_t) \\
 \text{Critic:} \quad V_\psi & \approx \mathbb{E}_{\pi_\phi, p_\theta} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \right]
 \end{aligned} \quad (2)$$

The overall loss of the agent can be found in Appendix C.

Methodology

We consider the visual task as a Partially Observable Markov Decision Process (POMDP) (Bellman 1957) due to the partial state observability from images. We denote the state space, the observation space, the action space and the reward function as \mathcal{S} , \mathcal{O} , \mathcal{A} and r respectively. The goal for an agent is to find a policy π^* to maximize the expected cumulative return $E_p(\sum_{t=1}^T r_t)$. As shown in Fig. 2, our method decouples the pretraining period from the downstream control task and thus consists of two main parts: Adaptive Pretraining and Policy Learning, which are described as follows.

Adaptive Pretraining

Dynamic adjustment on data augmentation compositions is applied on MoCo v2 to explore the importance of visual encoder in RL methods. Instead of providing a complete search

space for pretext task, APE provides the network with alternative compositions to learn robust and generalized representations. Specifically, two image features q and k^+ extracted from two augmented views of a same image serve as a query (He et al. 2020) and a key. The set $\{k^-\}$ is made up of the outputs from other images as negative samples. For each augmentation composition, InfoNCE (van den Oord, Li, and Vinyals 2018) is applied to maximize the agreement between q and k^+ :

$$\ell_q = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_{k^-} \exp(q \cdot k^- / \tau)} \quad (3)$$

where τ is a temperature parameter and all the embeddings are ℓ_2 normalized. In our augmentation strategy, each batch is divided into N sub-batches with the sampling probability p_i , i.e., $\sum_{i=1}^N p_i = 1$, which is initialized as $1/N$ for a fair assignment. The overall loss \mathcal{L}_z of all the augmentation compositions is formulated as follows:

$$\mathcal{L}_z = \sum_{i=1}^n \ell_q p_i \quad (4)$$

Here \mathcal{L}_z enables the encoder networks to maintain consistency across all sub-batches by utilizing the same key and query encoder. The closed-loop feedback structure works by utilizing the sampling probability, which is dynamically updated at the end of every epoch by:

$$p^{t+1} = \text{Softmax}(\alpha(1 - \text{Acc}^t)) \quad (5)$$

where α is set to 0.8 for 7 compositions, and 1 for 3 compositions, thus speeds up the process of exploration when given more augmentation choices. This updating strategy decreases the size of those well-explored compositions and attaches more importance to the ones with lower pretext task accuracy in the next epoch.

Policy Learning

The pretrained encoder projects the high-dimensional image observations o_t into low-dimensional latent features z_t , which are then transferred to the downstream agents that learn a control policy. The first three layers of the encoder are frozen to maintain generalization ability while parameters in the last layer are optimized together with the world model to adapt to environments with distribution shifts.

All model parameters θ in the latent dynamics except for the frozen ones in visual encoder’s first three layers are optimized end-to-end to minimize the following objectives:

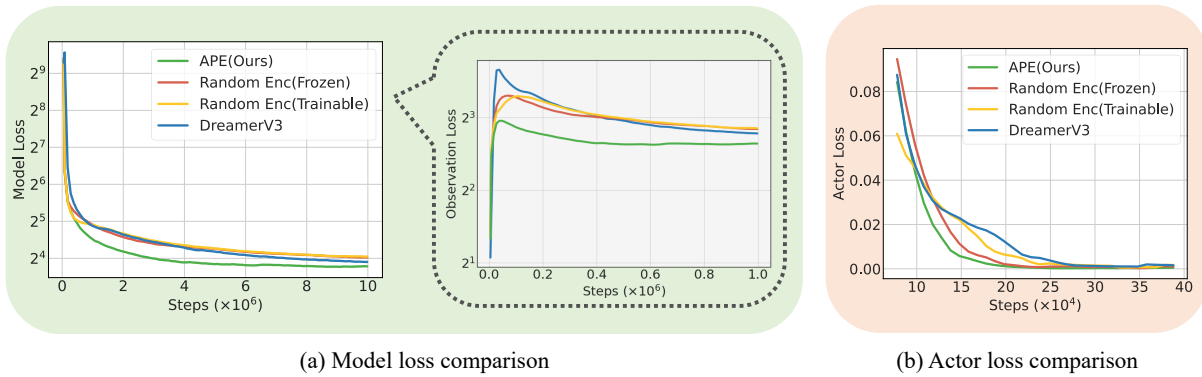


Figure 4: Loss comparison between DreamerV3, encoder with frozen random initialized parameters, encoder with trainable random initialized parameters and APE. The last layer of the frozen random initialized encoder is finetuned during training. The absolute value of actor loss is used.

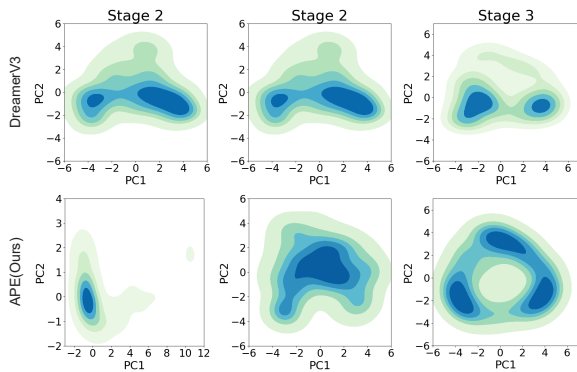


Figure 5: Exploration of states space in different phases during policy learning period. Data for 100 environment phases steps are sampled and visualized by Principal Component Analysis (PCA) in each stage. To compare fairly, axes are set to have identical ranges within the same stage. Thus the larger the state area, the higher the efficiency in exploration.

$$\begin{aligned}
 \mathcal{L}_{rew}(\theta) &= -\log(p_{\theta}^R(\hat{r}_t | z_t, z_{t-1}, a_{t-1})) \\
 \mathcal{L}_{con}(\theta) &= -\log(p_{\theta}^C(\hat{c}_t | z_t, z_{t-1}, a_{t-1})) \\
 \mathcal{L}_{rec}(\theta) &= -\log(g_{\theta}(\hat{x}_t | z_t, z_{t-1}, a_{t-1})) \\
 \mathcal{L}_{obs}(\theta) &= \beta_1 \max(1, \text{KL}[\text{sg}(f_{\theta}(z_t | z_{t-1}, a_{t-1}, o_t)) \\
 &\quad \| p_{\theta}^D(\hat{z}_t | z_{t-1}, a_{t-1})]) \\
 &\quad + \beta_2 \max(1, \text{KL}[(f_{\theta}(z_t | z_{t-1}, a_{t-1}, o_t)) \\
 &\quad \| \text{sg}(p_{\theta}^D(\hat{z}_t | z_{t-1}, a_{t-1}))])
 \end{aligned} \quad (6)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The fixed hyperparameters are set to $\beta_1 = 0.5$ and $\beta_2 = 0.1$. The overall loss of world model can be formulated as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{rew}(\theta) + \mathcal{L}_{con}(\theta) + \mathcal{L}_{rec}(\theta) + \mathcal{L}_{obs}(\theta) \quad (7)$$

Taking a multi-task view, the optimization of latent dynamics can be mainly divided into two parts, namely observation modeling and reward modeling (Ma et al. 2023).

Method	f_{main}	Acc. (%)
MoCo v2	—	90.84
APE	Jitter	91.08
APE	Blur	91.7

Table 1: Comparison of different augmentation settings using linear probes on ImageNet-100 validation set. We report top-5 classification accuracy and bold the highest result.

APE works by contributing to the first modeling task, which is attached more importance in MBRL frameworks.

Experiments

Several experiments are conducted to evaluate the performance of APE using fixed hyperparameters, with details provided in Appendix A and C. We investigate the following questions: (a) Can APE improve the agent’s generalization ability and sampling efficiency on various visual RL benchmarks? (b) Can APE generalize to both model-based and model-free methods? (c) Why APE works and how do choice of different settings affects the performance? By default, the encoder uses the ResNet18 architecture (He et al. 2015). Results reported are averaged over at least 3 runs.

Pretraining Encoders

We pretrain APE on ImageNet-100, a randomly selected subset of the common ImageNet-1k (Deng et al. 2009), which has also been utilized in pervious works (Kalantidis et al. 2020; Zhang, Zhu, and Yu 2023) for pretext tasks. Dynamic adjustment is made on the applied frequency of five data augmentations, including random color jittering, random grayscale conversion, random gaussian blur, random resized crop and random horizontal flip. Results under linear classification protocol are reported in Table 1. The augmentation with varying applied frequency during pretraining is denoted as the main augmentation strategy (f_{main}). In our method, the default f_{main} is random gaussian blur, which proved to be the most promising setting in AdDA.

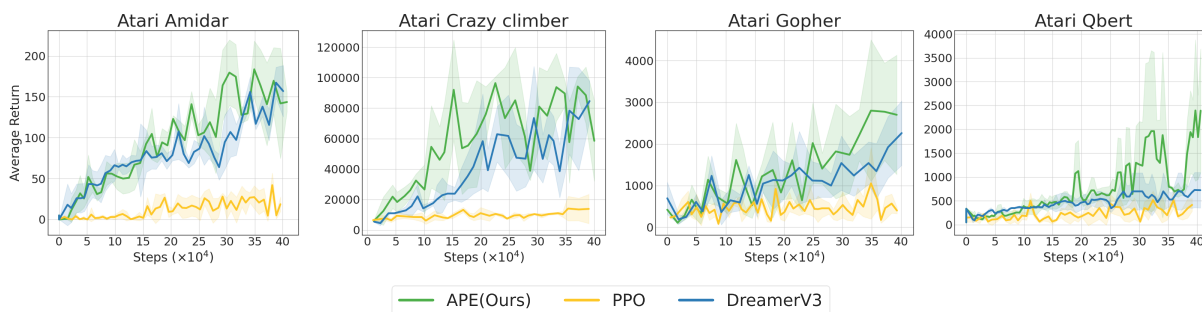


Figure 6: Training curves for Atari 100k benchmarks.

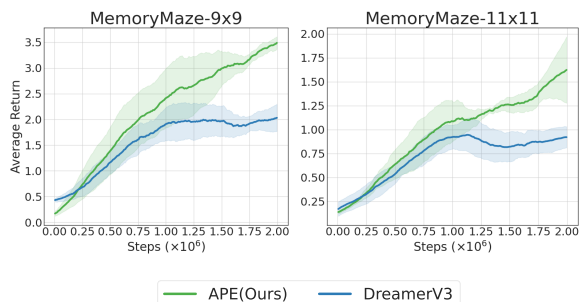


Figure 7: Training curves for Memory Maze benchmarks.

DMC Results

Being a widely studied benchmark in control tasks, DMC provides a reasonably challenging and diverse set of environments. We evaluate the sample efficiency of APE on DMC vision tasks for 1M environment steps. As shown in Fig.3, experiments conducted on those tasks demonstrates that APE benefits from the strong feature extraction capabilities learned from ImageNet, leading to enhanced training efficiency and asymptotic performance when applied to control tasks. Detailed comparison results of DMC scores are reported in Appendix B.

We illustrate the corresponding loss curves of DMC walker walk task learned with different encoders in Fig. 4. Encoders with random initialization have the same network architecture as APE, with frozen or trainable random initialized parameters. Intuitively, a pretrained encoder helps accelerate the convergence of observation loss (shown in Fig. 4(a)), since it provides prior knowledge for extracting visual features. Moreover, model loss demonstrated in Fig. 4(a) indicates that APE also helps in the multi-task optimization of latent dynamics, as the overall model loss with pretrained encoder converges more easily than others. Besides, actor loss in Fig. 4(b) suggests that world model equipped with improved encoder is able to predict better future outcomes of potential actions, and thus speed up the actor’s learning process. Furthermore, by visualizing the states space in Fig. 5, we demonstrate that APE enables more sufficient exploration in states with larger visualization area, thereby enhancing the downstream performance. Visualization of reconstructions is provided in Appendix B.

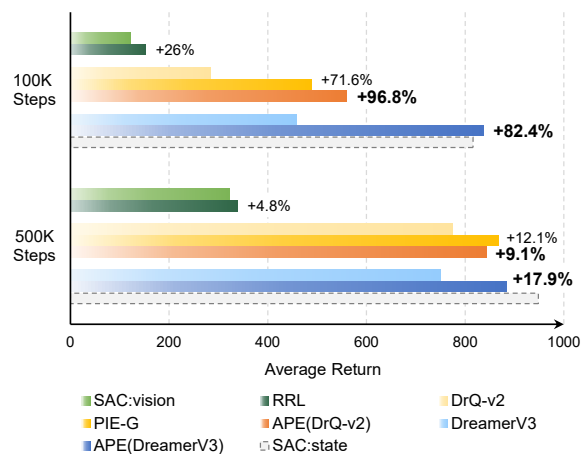


Figure 8: Comparison of DreamerV3-based and DrQ-v2-based APE against other ResNet pretrained algorithms (RRL and PIE-G), together with SAC:state, which learns on proprioceptive observations. The bars sharing the same color family (green, orange, and blue) denote algorithm groups following the same downstream strategy. The performance gains are calculated based on the RL backbone of each group (SAC:vision, DrQ-v2 and DreamerV3), with APE showing the most significant improvement.

Results on Other Benchmarks

Fig. 6 indicates that APE achieves better or comparable performance using same hyperparameters on 4 Atari tasks. This environment is often used as a benchmark for investigating data-efficiency in RL algorithms. Following the common setup of Atari 100k, we set the environment steps to 40k in tasks considered. The performance on Atari benchmarks highlights the robustness and generalization capability of APE in various RL settings.

Additional experiments have also been made on Memory Maze, which is a 3D domain of randomized mazes generated from a first-person perspective, which measures the long-term memory of the agent and requires it to localize itself by integrating information over time. In this paper, tasks on Memory Maze are trained for 2M steps due to limited computational resources. As shown in Fig. 7, APE is superior over the DreamerV3 baseline on these tasks that re-

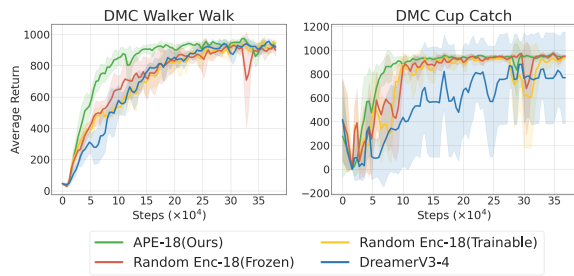


Figure 9: Choosing suitable pretraining strategy weighs more than increasing the depth of encoder network. We compare APE with random initialized encoder with frozen parameters, random initialized encoder with trainable parameters and DreamerV3. The last layer of the frozen random initialized encoder is finetuned during training. ‘-18’ and ‘-4’ denote the number of layers used in the encoder.

quire semantic understanding of the environment, making it a promising candidate for complex real-world applications requiring sophisticated decision-making processes.

Comparison with Other Pretrained Algorithms

As shown in Fig. 8, we compare the performances of two ResNet pretrained algorithms (RRL and PIE-G) and their base algorithms (SAC:vision and DrQ-v2) on three DMC benchmarks. APE outperforms all those methods on the 100K and 500K environment step benchmarks and achieves comparable performance with SAC:state (an agent that learns directly from states) at 100K environment step. To compare fairly, we reimplement DrQ-v2-based APE (denoted as APE (DrQ-v2)) to show that our findings and approach are not limited to MBRL framework. The results of SAC:state, RRL and DrQ-v2 are from the paper of RRL (Shah and Kumar 2021) and DrQ-v2 (Yarats et al. 2021a), while the others are reproduced and averaged over at least 3 runs. Detailed results are reported in the Appendix B.

Ablation Studies

Pretraining does work. Experiments are conducted to figure out whether deeper encoder helps to extract more discriminative features (shown in Fig. 9). *Random Encs* with frozen or trainable initialized parameters have the same network architecture as APE and are included as baselines to eliminate the effect of varied network size. By comparing the performance of *Random Enc* and DreamerV3, it is important to note that deeper networks do not always guarantee the extraction of better features, which leads to improved performance of APE in our tasks. This underscores the significant role of the pretraining period for RL algorithms.

Augmentations matter. In Fig. 10, we focus on the applied frequency of random gaussian blur and random color jittering to investigate the effect of data agumentations on visual representations in RL tasks. We observe that the sampling efficiency varies when changing the augmentation strategy. Results also indicates that linear probes may serve as a useful metric of pretrained model quality under the same

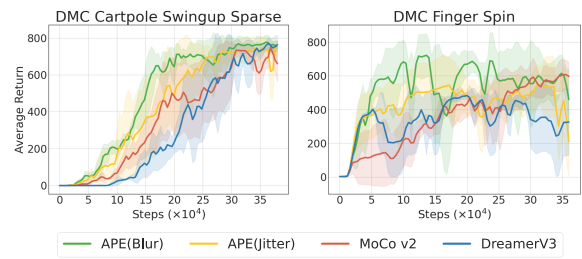


Figure 10: Different choices of augmentation strategy. APE with random gaussian blur as its main augmentation strategy outperforms other settings.

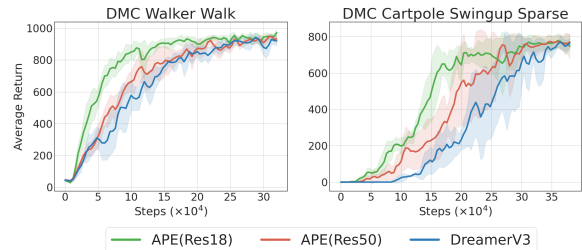


Figure 11: Different choices of network architectures. This figure indicates that APE with ResNet18 achieves better results compared with a deeper APE (ResNet50).

network architecture, with relative findings made on imitation learning (Hu et al. 2023).

Different choices of architectures. As shown in Fig. 11, we further explore the impact of architectures in DMC tasks. Following the settings of ResNet18 architecture, we freeze the first three layers of ResNet50 and update the last layer during training. For a fair comparison, the latent dimension of the three architectures are kept same (4096) and both the ResNet18 and the ResNet50 architecture are pretrained on ImageNet-100 with the same f_{main} . Results demonstrate that the increase of depth and complexity of the network, which lead to more abstract representations, may compromises the performance of the fine-grained control tasks. Comparisons with ViT-based pretrained encoder (He et al. 2021) are reported in the Appendix B.

Conclusion

In this paper, we propose APE, a simple yet effective method that implements adaptively pretrained encoder in RL frameworks. Unlike previous methods, APE is pretrained on a wide range of existing real-world images using a dynamic augmentation strategy, which helps the network to acquire more generalizable features in the downstream policy learning period. Experimental results show that our method surpasses state-of-the-art visual RL algorithms in learning efficiency and performance across various challenging domains. Besides, APE approaches the performance of state-based SAC in several control tasks, underscoring the effectiveness of augmentation strategy in the pretraining period.

Acknowledgments

This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS)(XDB1010302), CAS Project for Young Scientists in Basic Research, Grant No. YSBR-041 and the International Partnership Program of the Chinese Academy of Sciences (CAS) (173211KYSB20200021).

References

- Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132*.
- Baevski, A.; Zhou, H.; rahman Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv*, abs/2006.11477.
- Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2012. The Arcade Learning Environment: An Evaluation Platform for General Agents. *ArXiv*, abs/1207.4708.
- Bellman, R. 1957. A Markovian Decision Process. *Indiana University Mathematics Journal*, 6: 679–684.
- Brown, N.; Bakhtin, A.; Lerer, A.; and Gong, Q. 2020. Combining Deep Reinforcement Learning and Search for Imperfect-Information Games. *ArXiv*, abs/2007.13544.
- Burns, K.; Witzel, Z.; Hamid, J. I.; Yu, T.; Finn, C.; and Hausman, K. 2023. What Makes Pre-Trained Visual Representations Successful for Robust Manipulation? *ArXiv*, abs/2312.12444.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv*, abs/2002.05709.
- Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Delfosse, Q.; Stammer, W.; Rothenbacher, T.; Vittal, D.; and Kersting, K. 2022. Boosting Object Representation Learning via Motion and Object Continuity. *ArXiv*, abs/2211.09771.
- Delfosse, Q.; Sztwiertnia, S.; Stammer, W.; Rothmel, M.; and Kersting, K. 2024. Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents. *ArXiv*, abs/2401.05821.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.
- Du, Y.; Gan, C.; and Isola, P. 2021. Curious Representation Learning for Embodied Intelligence. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10388–10397.
- Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. *ArXiv*, abs/2106.07411.
- Gelada, C.; Kumar, S.; Buckman, J.; Nachum, O.; and Bellemare, M. G. 2019. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. *arXiv e-prints*, arXiv:1906.02736.
- Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
- Ha, D.; and Schmidhuber, J. 2018. World Models. *arXiv e-prints*, arXiv:1803.10122.
- Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic Algorithms and Applications. *ArXiv*, abs/1812.05905.
- Hafner, D.; Lillicrap, T. P.; Ba, J.; and Norouzi, M. 2019. Dream to Control: Learning Behaviors by Latent Imagination. *ArXiv*, abs/1912.01603.
- Hafner, D.; Lillicrap, T. P.; Norouzi, M.; and Ba, J. 2020. Mastering Atari with Discrete World Models. *ArXiv*, abs/2010.02193.
- Hafner, D.; Paukonis, J.; Ba, J.; and Lillicrap, T. P. 2023. Mastering Diverse Domains through World Models. *ArXiv*, abs/2301.04104.
- Hansen, N.; Su, H.; and Wang, X. 2021. Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation. In *Neural Information Processing Systems*.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Doll’ar, P.; and Girshick, R. B. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, 4182–4192. PMLR.
- Hu, Y.; Wang, R.; Li, L.; and Gao, Y. 2023. For Pre-Trained Vision Models in Motor Control, Not All Policy Learning Methods are Created Equal. In *International Conference on Machine Learning*.
- Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*, 30: 5875–5888.
- Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive

- learning. *Advances in Neural Information Processing Systems*, 33: 21798–21809.
- Kostrikov, I.; Yarats, D.; and Fergus, R. 2020. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. *ArXiv*, abs/2004.13649.
- Laskin, M.; Lee, K.; Stooke, A.; Pinto, L.; Abbeel, P.; and Srinivas, A. 2020. Reinforcement Learning with Augmented Data. *ArXiv*, abs/2004.14990.
- Lee, K.; Lee, K.; Shin, J.; and Lee, H. 2019. Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning. In *International Conference on Learning Representations*.
- Lin, Y.-C.; Zeng, A.; Song, S.; Isola, P.; and Lin, T.-Y. 2020a. Learning to See before Learning to Act: Visual Pre-training for Manipulation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 7286–7293.
- Lin, Z.; Wu, Y.-F.; Peri, S.; Sun, W.; Singh, G.; Deng, F.; Jiang, J.; and Ahn, S. 2020b. SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition. *ArXiv*, abs/2001.02407.
- Liu, P.; Wang, L.; Ranjan, R.; He, G.; and Zhao, L. 2022. A Survey on Active Deep Learning: From Model Driven to Data Driven. *ACM Comput. Surv.*, 54(10s).
- Liu, Q.; Zhou, Q.; Yang, R.; and Wang, J. 2023. Robust Representation Learning by Clustering with Bisimulation Metrics for Visual Reinforcement Learning with Distractions. In *AAAI Conference on Artificial Intelligence*.
- Ma, H.; Wu, J.; Feng, N.; Wang, J.; and Long, M. 2023. HarmonyDream: Task Harmonization Inside World Models.
- Ma, Y. J.; Sodhani, S.; Jayaraman, D.; Bastani, O.; Kumar, V.; and Zhang, A. 2022. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. *ArXiv*, abs/2210.00030.
- Paukonis, J.; Lillicrap, T. P.; and Hafner, D. 2022. Evaluating Long-Term Memory in 3D Mazes. *ArXiv*, abs/2210.13383.
- Poudel, R. P. K.; Pandya, H.; Liwicki, S.; and Cipolla, R. 2023. ReCoRe: Regularized Contrastive Representation Learning of World Model. *ArXiv*, abs/2312.09056.
- Raileanu, R.; Goldstein, M.; Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Automatic Data Augmentation for Generalization in Reinforcement Learning. In *Neural Information Processing Systems*.
- Schmid, M.; Moravčík, M.; Burch, N.; Kadlec, R.; Davidson, J.; Waugh, K.; Bard, N.; Timbers, F.; Lanctot, M.; Holland, Z.; Davoodi, E.; Christianson, A.; and Bowling, M. H. 2021. Player of Games. *ArXiv*, abs/2112.03178.
- Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T. P.; and Silver, D. 2019. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604 – 609.
- Schwarzer, M.; Rajkumar, N.; Noukhovitch, M.; Anand, A.; Charlin, L.; Hjelm, D.; Bachman, P.; and Courville, A. C. 2021. Pretraining Representations for Data-Efficient Reinforcement Learning. In *Neural Information Processing Systems*.
- Shah, R.; and Kumar, V. 2021. RRL: Resnet as representation for Reinforcement Learning. *ArXiv*, abs/2107.03380.
- Song, X.; Jiang, Y.; Tu, S.; Du, Y.; and Neyshabur, B. 2019. Observational Overfitting in Reinforcement Learning. *ArXiv*, abs/1912.02975.
- Srinivas, A.; Laskin, M.; and Abbeel, P. 2020. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. *ArXiv*, abs/2004.04136.
- Stooke, A.; Lee, K.; Abbeel, P.; and Laskin, M. 2020. Decoupling Representation Learning from Reinforcement Learning. *ArXiv*, abs/2009.08319.
- 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.; and Riedmiller, M. A. 2018. DeepMind Control Suite. *ArXiv*, abs/1801.00690.
- van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv*, abs/1807.03748.
- Wang, K.; Kang, B.; Shao, J.; and Feng, J. 2020. Improving Generalization in Reinforcement Learning with Mixture Regularization. *ArXiv*, abs/2010.10814.
- Xiao, T.; Radosavovic, I.; Darrell, T.; and Malik, J. 2022. Masked Visual Pre-training for Motor Control. *ArXiv*, abs/2203.06173.
- Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021a. Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning. *ArXiv*, abs/2107.09645.
- Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021b. Reinforcement Learning with Prototypical Representations. In *International Conference on Machine Learning*.
- Yarats, D.; Zhang, A.; Kostrikov, I.; Amos, B.; Pineau, J.; and Fergus, R. 2019. Improving Sample Efficiency in Model-Free Reinforcement Learning from Images. In *AAAI Conference on Artificial Intelligence*.
- Yu, T.; Zhang, Z.; Lan, C.; Chen, Z.; and Lu, Y. 2022. Mask-based Latent Reconstruction for Reinforcement Learning. *ArXiv*, abs/2201.12096.
- Yuan, Z.; Xue, Z.; Yuan, B.; Wang, X.; Wu, Y.; Gao, Y.; and Xu, H. 2022. Pre-Trained Image Encoder for Generalizable Visual Reinforcement Learning. *ArXiv*, abs/2212.08860.
- Zhan, A.; Zhao, P.; Pinto, L.; Abbeel, P.; and Laskin, M. 2020. Learning Visual Robotic Control Efficiently with Contrastive Pre-training and Data Augmentation. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4040–4047.
- Zhang, Y.-H.; Zhu, H.; and Yu, S. 2023. Adaptive Data Augmentation for Contrastive Learning. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.
- Zhao, Y.; Wu, K.; Yi, T.; Xu, Z.; Ju, X.; Che, Z.; Qiu, Q.; Liu, C. H.; and Tang, J. 2024. An Efficient Generalizable Framework for Visuomotor Policies via Control-aware Augmentation and Privilege-guided Distillation. *ArXiv*, abs/2401.09258.