International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02472-w

A Comprehensive Survey of Data Augmentation in Visual Reinforcement Learning

Guozheng Ma[1] · Zhen Wang[2] · Zhecheng Yuan[3] · Xueqian Wang[4] · Bo Yuan[5] · Dacheng Tao[1]

Received: 1 December 2022 / Accepted: 3 May 2025

Abstract
Visual reinforcement learning (RL), which makes decisions directly from high-dimensional visual inputs, has demonstrated significant potential in various domains. However, deploying visual RL techniques in the real world remains challenging due to their low sample efficiency and large generalization gaps. To tackle these obstacles, data augmentation (DA) has become a widely used technique in visual RL for acquiring sample-efficient and generalizable policies by diversifying the training data. This survey aims to provide a timely and essential review of DA techniques in visual RL in recognition of the thriving development in this field. In particular, we propose a unified framework for analyzing visual RL and understanding the role of DA in it. We then present a principled taxonomy of the existing augmentation techniques used in visual RL and conduct an in-depth discussion on how to better leverage augmented data in various scenarios. Moreover, we report the empirical evaluation of DA-based techniques in visual RL and conclude by highlighting the directions for future research. As the first comprehensive survey of DA in visual RL, this work is expected to offer valuable guidance to this emerging field.

Keywords Visual Reinforcement Learning · Representation Learning · Data Augmentation · Regularization.

1 Introduction

Reinforcement learning (RL) addresses sequential decision-making problems in which an agent seeks to discover the optimal policy via trial-and-error interactions with the environment [40, 100, 120, 143]. Visual RL, a variant that learns directly from visual observations such as images, has gained widespread application across various domains due to its intuitive and cost-effective approach to environmental perception [170, 192]. This paradigm has been successfully employed in video games [151], autonomous driving [82], robot control [76], and other areas. However, learning directly from high-dimensional visual observations remains largely hindered by the challenges of low sample efficiency and large generalization gaps [55, 88, 189, 190, 192].

To learn sample-efficient and generalizable visual RL agents, a considerable amount of effort has been devoted to developing diverse approaches, including (1) applying **explicit regularization** techniques such as entropy regularization [48, 206] to constrain the model's weights [23, 58, 107]; (2) performing joint learning with RL loss and **auxiliary tasks** to provide additional representation supervision [4, 34, 68, 80, 89, 94, 117, 129, 144, 157, 192–195, 214]; (3) building a **world model** of the RL environment that allows learning behaviors from imagined outcomes [49–51, 92, 169]; and (4) **pretraining an encoder** that can project high-dimensional observations into compact state representation [91, 104, 137, 142, 147, 156, 166, 179, 191, 199].

Although these approaches have achieved remarkable success, they remain challenged by limited interaction data and poor sample diversity [88, 189, 190]. To address these lim-

Communicated by Gang Hua.

✉ Bo Yuan
boyuan@ieee.org

✉ Dacheng Tao
dacheng.tao@ntu.edu.sg

1 College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

2 School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, Australia

3 Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

4 Shenzhen International Graduate School, Tsinghua University, Beijing, China

5 School of Electrical Engineering and Computer Science, The University of Queensland, Queensland, Australia

Published online: 28 July 2025

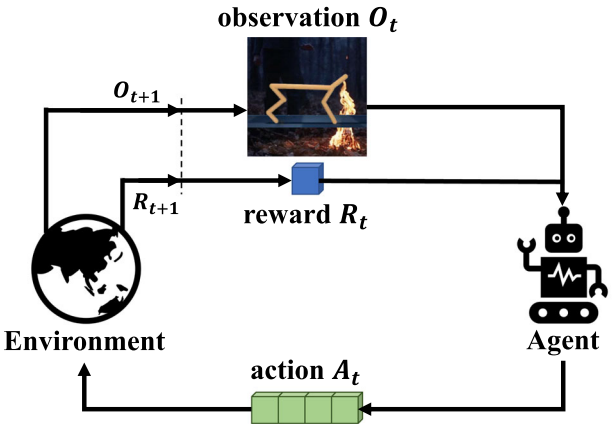


Fig. 1 Agent-environment interaction loop of visual RL: Direct decision-making from high-dimensional observations brings both flexibility and challenges.

itations by increasing the quantity and diversity of training data, data augmentation (DA) has garnered increasing attention from the visual RL community in recent years [35, 55, 190]. As a data-driven method, DA is orthogonal to the aforementioned approaches and can be combined with them to further enhance performance [144, 168]. For instance, DA plays a crucial role in contrastive-based auxiliary tasks, injecting prior knowledge of task invariance [69, 89, 156]. In addition, DA is essential for pre-training a cross-task representation [156, 191]. Furthermore, various DA techniques, such as random cropping, have been incorporated into almost all visual RL algorithms as a form of data preprocessing [51, 129, 170].

In general, DA refers to strategies for generating synthetic training data from existing data without additional collection or interaction efforts [38, 150]. Figure 2 illustrates the generic workflow for leveraging DA in visual RL: diverse augmented data are generated by manipulating the original interaction data and then exploited to optimize the RL objective [88, 189, 190]. Moreover, DA can further enhance the representation learning in visual RL by incorporating auxiliary objectives [54, 89, 141, 144]. Despite the surge of related studies on leveraging DA in visual RL scenarios, this fast-evolving and expanding field still lacks clarity and coherence. Therefore, this comprehensive survey aims to provide a bird's-eye view of DA-based methods in visual RL with the following main contributions:

1. Based on previous works [83, 154], we present High-Dimensional Contextual Markov Decision Process (HCMDP) as a general framework to formalize visual RL. This framework provides deep insights into the challenges of **low sample efficiency** and **large generalization gaps** in visual RL, which serve as the primary motivations for introducing DA.
2. We identify two key assumptions of DA with different motivations: the **optimality invariance** assumption for improving sample efficiency and the **prior-based diversity** assumption for narrowing the generalization gap.
3. We categorize related studies from two principled perspectives: **how to augment data** and **how to leverage augmented data** for improved clarity and coherence. This classification provides a structured framework for systematically reviewing existing work, offering a clear and logical organization of the field's current state.
4. We conduct **a systematic empirical evaluation** of extensive DA-based methods on representative benchmarks to comprehensively assess their performance in terms of sample efficiency and generalization capabilities.
5. We present **a comprehensive analysis** of DA in visual RL as the cornerstone of this survey, systematically examining its unique mechanisms, significant challenges, and emerging opportunities in the field. Through this in-depth analysis, we provide critical insights into both the current landscape and future prospects of DA in visual RL, along with detailed discussions on potential research directions and practical considerations.

The body of this survey is organized as Fig. 3. In Section 2, we propose a unified high-dimensional contextual

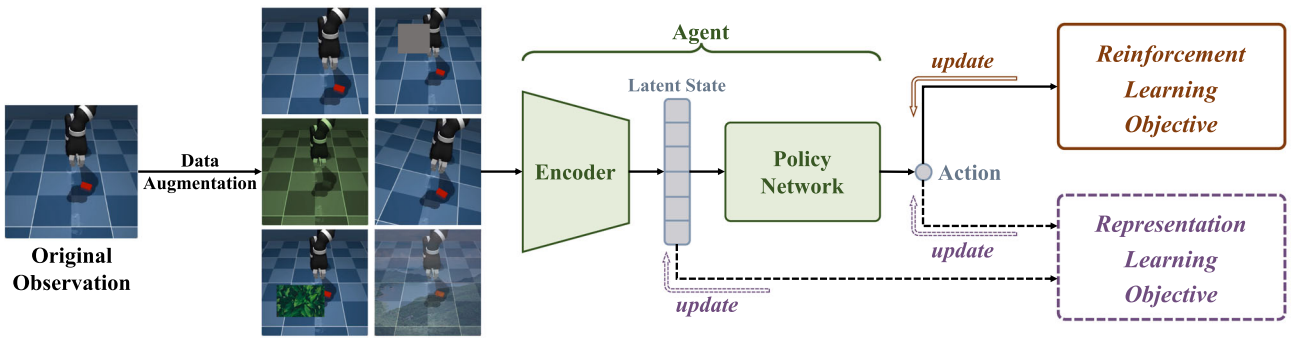


Fig. 2 The generic workflow diagram for leveraging DA in visual RL.

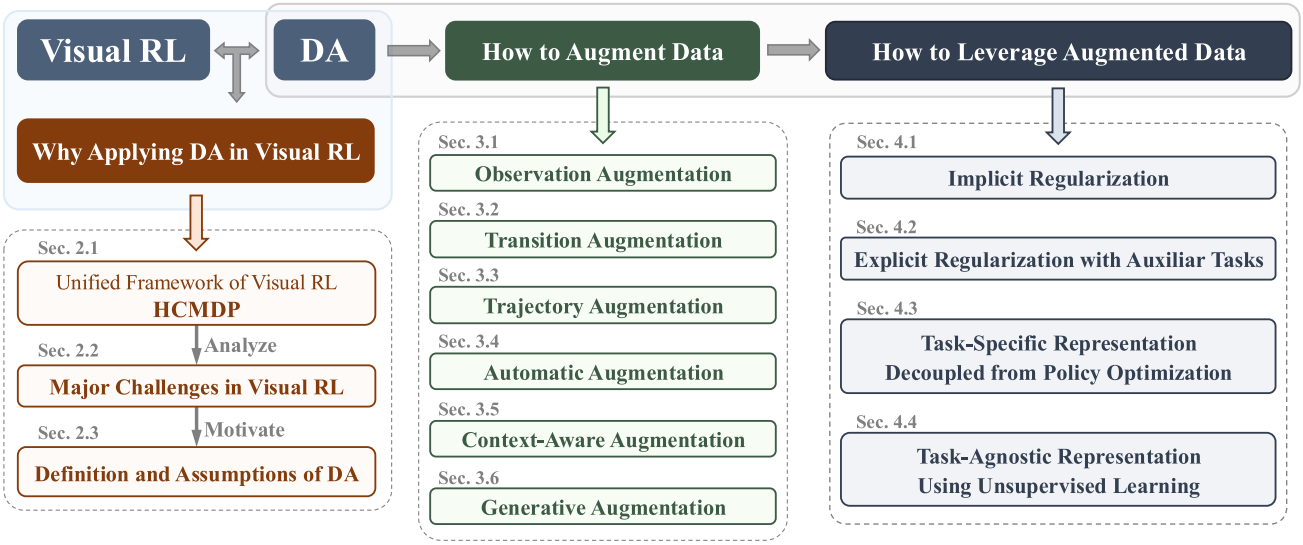


Fig. 3 The schematic structure of this survey.

Markov decision process (HCMDP) framework (Section 2.1) to formalize the visual RL scenario and highlight its major challenges (Section 2.2), as well as present the motivations and definitions of DA in visual RL (Section 2.3). We then conduct a systematic review of the previous work from two perspectives: how to obtain and how to leverage augmented data in visual RL (Section 3 and Section 4). In Section 3, we categorize DA approaches in visual RL into *observation augmentation*, *transition augmentation*, and *trajectory augmentation*, based on the type of data each technique aims to modify. Moreover, we introduce three advanced extensions: *automatic augmentation*, *context-aware augmentation*, and *generative augmentation*. In Section 4, we present the different mechanisms used to leverage DA in visual RL, including implicit and explicit regularization, task-specific representation learning decoupled from policy optimization, and task-agnostic representation learning using unsupervised learning. To reveal the practical effect of DA, we introduce the typical benchmarks and summarize the empirical performance of recent DA-based methods in Section 5. In Section 6, we put forward a critical discussion concerning future research directions, including the opportunities, challenges, limitations, and underlying mechanisms of DA. Finally, this survey is concluded in Section 7 with a list of key insights.

Scope.

Given the multitude of topics and research areas related to DA and visual RL, we constrain the scope of this survey in several ways to ensure its feasibility. *Firstly*, this survey does not cover the related topic of domain randomization (DR) [138, 162], which aims to solve the sim-to-real problem in robot control by tuning the physical simulator's parameter distribution to align as closely as possible with reality [66, 71, 125]. In contrast, DA can only manipulate observations post-rendering, without access to the simulator's internal parameters, which affords it greater flexibility [83]. *Secondly*, this survey focuses on scenarios that involve learning directly from visual inputs (visual RL) rather than handcrafted state inputs (state-based RL). Consequently, several works that introduce DA in state-based RL, will not be prominently featured in this survey [102, 108]. *Thirdly*, while DA is a powerful tool, it is not the sole approach for improving sample efficiency and generalization in visual RL. To maintain coherence and focus, this survey does not provide detailed coverage of works that use DA as a foundational technique but whose primary research focus lies elsewhere [146, 181, 200]. We strongly recommend readers interested in generalization issues in RL to refer to another comprehensive survey [83] that focuses on generalization in deep RL. *Finally*, to ensure this survey aligns with the latest developments in AI field, we provide a detailed introduction to recent research on DA using advanced generative models in Section 3.6. Additionally, in Section 6.5, we critically examine the role and relevance of visual RL and DA in the context of the rapidly evolving landscape of foundation models.

2 Preliminaries

Visual RL addresses high-dimensional image observations instead of well-designed states and has encountered a series of new challenges [189, 192]. This section analyzes visual RL in depth and introduces the formalism of DA used for visual RL. In Section 2.1, we present a novel framework, HCMDP, to formalize the paradigm of visual RL. Based on this framework, we analyze the major challenges faced by visual RL

in Section 2.2. Finally, Section 2.3 introduces the formalism of DA in visual RL, including its motivation, definition and two key assumptions.

2.1 High-Dimensional Contextual MDP (HCMDP)

The standard RL task is often defined as a **Markov Decision Process (MDP)** [120], which is specified by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, p, \gamma)$ where $\mathcal{S}$ is the state space; $\mathcal{A}$ is the action space; $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is the scalar reward function; $\mathcal{P}(s'|s, a)$ is the transition function; $p(\cdot)$ is the initial state distribution; and $\gamma \in (0, 1]$ is the discount factor. The goal of RL is to learn an optimal policy $\pi^*(a|s)$ that maximizes the expected cumulative discounted return $\mathcal{R}(\pi, \mathcal{M})$, which is defined as:

$$\mathcal{R}(\pi, \mathcal{M}) = \mathbb{E}_{\substack{s_0 \sim p(\cdot) \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \right] \tag{1}$$

Although the MDP is the standard paradigm of RL, it ignores a crucial factor of visual RL: agents only have direct access to high-dimensional observations instead of the actual state information. To properly formulate the visual RL scenarios, as shown in Fig. 1, many variants of MDPs [31, 32, 41, 53] have been introduced by using the high-dimensional observation space $\mathcal{O}$ to represent the image inputs. Depending on the specific assumptions, an emission function $\phi : \mathcal{S} \mapsto \mathcal{O}$ can be designed to simulate the mapping from the state space $\mathcal{S}$ to the observation space $\mathcal{O}$. For example, the (f, g)-scheme [154] constructs an emission function as the combination of generalizable and non-generalizable features while the contextual MDP (CMDP) [31, 52, 83] introduces context c to distinguish contextual information from the underlying state information. However, these MDP variants mainly focus on how to explain the generalization effect in visual RL, and ignore the issue of constructing a compact representation from high-dimensional observations.

To better understand visual RL scenarios and provide a unified view of its specific challenges, we propose **High-Dimensional Contextual MDP (HCMDP)** as a general modeling framework of visual RL. Following the previous formalism [83, 154], the HCMDP $\mathcal{M}|_C$ can be defined as a family of environments:

$$\mathcal{M}|_C = \{\mathcal{M}|_c = (\mathcal{M}, \mathcal{O}_c, \phi_c) | c \sim p(c), c \in C\} \tag{2}$$

where $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, p, \gamma)$ specifies the dynamics of the underlying system. With the fixed base MDP $\mathcal{M}$, the observation space $\mathcal{O}_c$ and emission function ϕ_c depend on the context c, which refers to the peripheral parameters that are not essential for agents to make decisions. $p(c)$ is the *context distribution*, and C represents the entire *context set*. For example, the colors and styles of backgrounds in robot scenarios are extraneous to control tasks, and are thus being referred to as task-irrelevant features.

To be more specific, context c can be denoted as a set of parameters $\{c_1, c_2, \ldots, c_n\}$, where n is the number of task-irrelevant properties in this system. Each c_i corresponds to a task-irrelevant property, all of which are distributed over a fixed range: $\{c_1 \in C_1, c_2 \in C_2, \ldots, c_n \in C_n\}$. Consider an autonomous driving example such as CARLA [215]: an agent learns to control the car directly from pixels in changing environments. Therefore, the agent must distinguish between task-relevant and task-irrelevant components in the image observations. For instance, we can denote the style of the background buildings as c_1, the color of the driving car as c_2 and the number of people walking on the sides of the road as c_3.

The state s and context c constitute the complete information (parameters) used by the system to render the final observed images [154]. However, they both exist in the low-dimensional latent space, which cannot be directly observed. In fact, $\mathcal{O}$ is the only observable high-dimensional space where agents perceive task information. Following the assumptions [154, 163] that observations are high-dimensional projections of the state s and task-irrelevant contexts c, the emission function ϕ_c mapping from state $s \in \mathcal{S}$ to observation $o \in \mathcal{O}_c$ can be defined as:

$$o = \phi_c(s) := \mathbf{h}(s^H, c_1^H, c_2^H, \ldots, c_n^H) \tag{3}$$

where s^H is the high-dimensional representation mapped from the underlying state s, and each c_i^H is the representation uniquely determined by the latent context c_i. Similar to the formalism in [154], $\mathbf{h}$ is a "combination" function that combines the task-relevant state representation s^H and task-irrelevant context representations $(c_1^H, c_2^H, \ldots, c_n^H)$ to render the final observation. Based on the HCMDP framework, Fig. 4 shows an illustration of a robot control environment from the DeepMind control suite [161]. In this scenario, contexts c_1 and c_2 separately denote the floor color and background style, respectively, which are both irrelevant to the control task. Correspondingly, c_1^H and c_2^H are the high-dimensional representations mapped from c_1 and c_2. The final observation o is the combination of the state representation s^H and the task-irrelevant representations c_1^H and c_2^H.

An HCMDP $\mathcal{M}|_C$ consists of a family of specific environments, where c follows the *context distribution* $p(c)$ over the entire *context set* C. In a given system, $\mathcal{M}$ and the rendering rules from s and c_i to the high-dimensional representations s^H and c_i^H are established. Hence, different combinations of the *context distribution* $p(c)$ and *context set* C produce different HCMDPs. For any HCMDP $\mathcal{M}|_C$, the expected return of a policy is defined as:

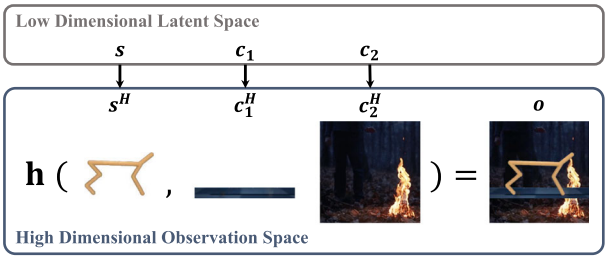


Fig. 4 A visualized example of HCMDP.

$$\mathbf{R}(\pi, \mathcal{M}|_C) := \mathbb{E}_{c\sim p(c), c\in C}\left[\mathcal{R}\left(\pi, \mathcal{M}|_c\right)\right] \tag{4}$$

where $\mathcal{R}$ is the expected return of policy π in a specific MDP. In practice, we assume that the context distribution is uniform over the entire context set [83] so that different HCMDPs can be specified by their context sets $C = (C_1, C_2, \ldots, C_n)$. By choosing a training context set C_{train} and a test context set C_{test}, we can separately define the training context set HCMDP $\mathcal{M}|_{C_{\text{train}}}$ and the test context set HCMDP $\mathcal{M}|_{C_{\text{test}}}$. Agents are only allowed to be trained in $\mathcal{M}|_{C_{\text{train}}}$ and evaluated in the same HCMDP $\mathcal{M}|_{C_{\text{train}}}$ or HCMDP $\mathcal{M}|_{C_{\text{test}}}$, whose context exhibits a distribution shift from the training context set.

Remarks. The key distinction between HCMDP and other MDP variants lies in the emission function $\phi : \mathcal{S} \mapsto \mathcal{O}$, as illustrated in Fig. 5.

First, HCMDP explicitly characterizes the high dimensionality of the observation space $\mathcal{O}$ by specifying the mapping between the latent variables s, c and their high-dimensional representations s^H, c^H. Second, it provides a unified perspective for understanding generalization challenges when deploying learned policies to unseen visual environments, building upon existing assumptions [154, 204]. Specifically, HCMDP posits that the task-relevant features of state s and task-irrelevant features of context c are combined in the final observation without imposing additional assumptions about their relationship. During training, this leads agents to potentially overfit to irrelevant context features, hindering effective generalization to unseen environments.

As a general framework, HCMDP can be specialized into other MDP variants through additional assumptions. For instance, the (f, g)-scheme [154] assumes that non-generalizable features in observations are projected from the latent state via a function $g_\theta(\cdot)$ parameterized by θ; Block MDP [204] models the emission function as a concatenation of state variables and spurious noise: $s \oplus f(\eta)$; and BC-MDP [152] restricts the agent's access to a context-dependent partial state space $\mathcal{S}^c$. While these variants make specific assumptions about feature relationships, HCMDP takes a more general approach by focusing solely on how

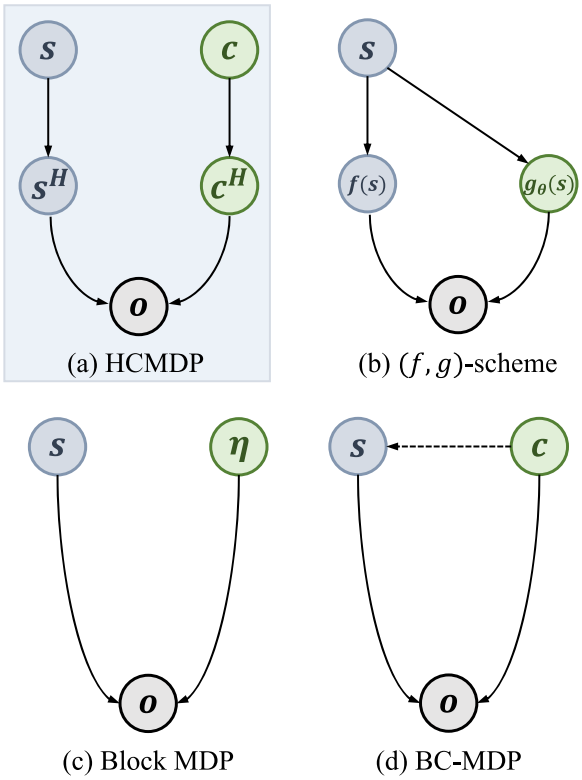


Fig. 5 A graphical model of the emission function of HCMDP (a) compared with three other representative MDP variants: (b) (f, g)-scheme [154], (c) Block MDP [32, 204], and (d) BC-MDP [152].

task-relevant and task-irrelevant features compose the final observation.

Note that the HCMDP framework does not take into account the partially observable features of the underlying states in a partially observable MDP (POMDP) [57]. Following [120, 189, 191], we assume that the complete state information can be reasonably constructed by stacking three consecutive previous image observations into a trajectory snippet [190]. In summary, the motivation of HCMDP is to emphasize the fact that the underlying state s is projected to the high-dimensional observation space along with the task-irrelevant information of context c. With this unified framework, the unique challenges of visual RL scenarios compared with standard RL can be clearly analyzed.

2.2 Major Challenges in Visual RL

Despite the success of visual RL in complex control tasks with visual observations, sample efficiency and generalization remain two major challenges that may lead to ineffective agents [35, 55, 136, 144, 183]. In this subsection, we present the formal definitions of sample efficiency and the generalization gap based on the HCMDP framework and discuss their mechanisms.

2.2.1 Sample Efficiency

This term measures how well the interaction data are leveraged to train a model [197]. In practice, we consider an agent sample-efficient if it can achieve satisfactory performance within limited environment interactions [144, 189]. In other words, the goal of sample-efficient RL is to maximize the policy's expected return during the training of HCMDP $\mathcal{M}|_{C_{\text{train}}}$ based on as few interactions as possible. The expected return of policy π in $\mathcal{M}|_{C_{\text{train}}}$ can be defined as:

$$J(\pi) := \mathbf{R}(\pi, \mathcal{M}|_{C_{\text{train}}}) \tag{5}$$

Instead of making decisions based on predefined features, the agent in visual RL need to learn an appropriate representation that maps a high-dimensional observation $\mathbf{h}(s^H, c^H)$ to the latent space $\mathbf{h}(s, c)$ to obtain decision-critical information [88, 144, 154, 189]. Since standard RL algorithms already require large amounts of interaction data [48], learning directly from high-dimensional observations suffers from prohibitive sample complexity [192].

One solution to the sample inefficiency problem in visual RL is by training with auxiliary losses, such as pixel or latent reconstruction [192, 195], future prediction [68, 94, 144, 194] and contrastive learning for instance discrimination [34, 80, 89, 157] or temporal discrimination [4, 117, 129, 135, 214]. Meanwhile, several model-based methods explicitly build a world model of the RL environment in pixel or latent spaces to conduct planning [49–51, 92]. Recently, pretrained encoders have demonstrated great potential in downstream tasks where the visual RL environment is explored in an unsupervised manner to obtain a task-agnostic pretrained encoder that can quickly adapt to diverse downstream tasks [91, 104, 156, 191]. In addition, applying the pretrained encoders from other domains such as ImageNet [27] to visual RL also has shown its efficiency in downstream tasks [137, 147, 166, 179]. The aforementioned methods have significantly improved the sample efficiency of visual RL, but the lack of training data remains a fundamental issue, which can be effectively solved by DA. Moreover, abundant auxiliary tasks and world models are designed and trained based on the augmented data [89, 92, 144, 194]. Hence, DA plays a vital role in improving the sample efficiency of visual RL algorithms.

2.2.2 Generalization

An agent's generalization ability can be measured by the generalization gap when transferred to unseen environments, which has been extensively investigated [32, 154, 163] and reviewed [83]. For an HCMDP with varying context sets C_{train} and C_{test}, the generalization gap of policy π can be defined as:

$$\text{GenGap}(\pi) := \mathbf{R}(\pi, \mathcal{M}|_{C_{\text{train}}}) - \mathbf{R}(\pi, \mathcal{M}|_{C_{\text{test}}}) \tag{6}$$

As mentioned in Section 2.1, the task-relevant information of state s is often conflated with the task-irrelevant information of context c, which may cause agents to overfit the task-irrelevant components [34, 154]. How to train generalizable agents across different environments remains challenging in visual RL, and distinguishing between the task-relevant and task-irrelevant components of the observed images is essential for narrowing the generalization gap.

A naive approach to enhancing generalization is to apply regularization techniques originally developed for supervised learning [23, 107], including ℓ_2 regularization [36], entropy regularization [48, 206], dropout [58] and batch normalization [72]. However, these traditional regularization techniques show limited improvement in generalization and may even negatively impact sample efficiency [23, 72, 189]. As a result, recent studies focus on learning robust representations to improve the agent's generalization ability by introducing bisimulation metrics [77, 205], multi-view information bottleneck (MIB) [34], pretrained image encoder [199] etc. From an orthogonal perspective, DA has been effective in enhancing generalization by generating diverse synthetic data [88, 189]. Moreover, DA can implicitly provide prior knowledge to the agent as a type of inductive bias or regularization [65, 83]. A detailed elaboration of the generalization issue in RL is provided in [83], which systematically reviews the related studies.

Remarks The primary purpose of establishing precise mathematical definitions within our HCMDP framework is to help readers formally comprehend the specific challenges in visual RL. Equation 5 and Equation 6 serve this purpose by mathematically formalizing sample efficiency and generalization challenges respectively. Beyond mere formalization, these equations serve several essential functions:

1. They precisely capture the fundamental challenge of visual RL - the entanglement of task-relevant and task-irrelevant information in high-dimensional observations - which distinguishes it from traditional state-based RL;
2. They transform intuitive concepts of sample efficiency and generalization into well-defined, quantifiable metrics, enabling rigorous analysis of these challenges;
3. They provide a framework for systematically examining and comparing how different methods, particularly DA, address these challenges through distinct mechanisms.

Together with the HCMDP modeling framework, these mathematical definitions form the foundation for understanding the unique challenges of visual RL and analyzing how different approaches, especially DA, address them.

2.3 DA in Visual RL

As discussed in Section 2.2, the quantity and diversity of training data are crucial for achieving sample-efficient and generalizable visual RL algorithms. DA, as a data-driven approach, has demonstrated significant potential for visual RL in terms of both sample efficiency and generalization ability [35, 55, 88, 141, 156, 189, 190, 198]. The advantages of DA for visual RL can be viewed from two aspects: (1) it can significantly expand the volume of the original interaction data, thus improving the sample efficiency [88]; (2) it introduces additional diversity into the original training data, making agents more robust to variations and enhancing their generalization capabilities [55, 83]. Recent studies have also revealed the regulatory effect of DA, which can accelerate the training process and prevent overfitting [65, 121]. Furthermore, theoretical foundations have also been developed for DA, such as invariance learning [14, 73] and feature manipulation [149]. Hence, DA has been well recognized as a viable solution for the challenges in visual RL [83, 185]. Following the conventions in [55, 190], we define a general augmentation $\tau : \mathcal{O} \times \mathcal{V} \mapsto \mathcal{O}^{aug}$ as a mapping from the original observation space $\mathcal{O}$ to the augmented observation space $\mathcal{O}^{aug}$:

$$o^{aug} \triangleq \tau(o; \nu) \qquad \forall o \in \mathcal{O}, \nu \in \mathcal{V} \tag{7}$$

where $\nu \in \mathcal{V}$ is a set of random parameters and $\tau(\cdot)$ is the transformation function acting on the observation o. To gain an intuitive understanding of the effect of DA, we identify two assumptions of $\tau(\cdot)$ corresponding to the challenges that DA seeks to address: the assumption of **optimality invariance** for improving the sample efficiency and the assumption of **prior-based diversity** for narrowing the generalization gap.

2.3.1 Optimality Invariance

In supervised learning (SL), DA methods usually assume that the model's output is invariant after transformations; therefore, they can be directly applied to labeled samples to produce supplementary data [29, 150]. Considering the property of RL, DrQ [189] defines the *optimality invariance* assumption as adding a constraint to the transformation τ, which induces an equivalence relation between state s and its augmented counterpart s^{aug} constructed from observations o and o^{aug}, respectively [55]. Hence, an optimality-invariant state transformation $\tau : \mathcal{O} \times \mathcal{V} \mapsto \mathcal{O}$ can be defined as a mapping that preserves the Q-values [55], V-values and policy π [141]:

$$\begin{aligned} &Q(o, a) = Q(\tau(o; \nu), a), \\ &V(o) = V(\tau(o; \nu)), \\ &\pi(o) = \pi(\tau(o; \nu)), \quad \forall o \in \mathcal{O}, a \in \mathcal{A}, \nu \in \mathcal{V} \end{aligned} \tag{8}$$

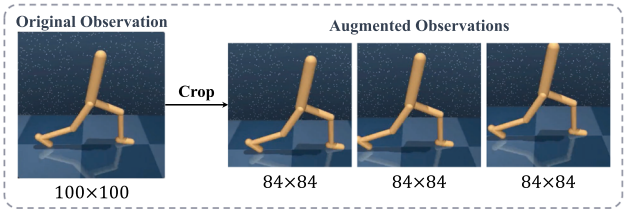


Fig. 6 Examples of the optimality-invariant augmentation, where key control-relevant information is preserved.

where ν is the set of parameters of $\tau(\cdot)$, drawn from the set of all possible parameters $\mathcal{V}$. Note that optimality invariance relies on strict restrictions on $\tau(\cdot)$ and the size of $\mathcal{V}$ to ensure that the same s can be constructed from the original and augmented observations. In the HCMDP framework, optimality invariance means that augmentation transformations only change the selected contexts in the high-dimensional observation space while preserving the entire (conceptual) state information in the latent space.

For instance, random cropping [88, 189] satisfies the optimality invariance assumption in most robot control environments such as the DeepMind control suite [161]. In Fig. 6, cropping generates augmented observations by randomly extracting central patches from the original image. Since the robot is centrally placed in the images, cropping only eliminates irrelevant information such as the background color while preserving the task-relevant information such as the robot's posture [163].

With the optimality-invariant augmentation of the original observations, we can obtain sufficient training data based on limited interactions with the environment so that the sample efficiency can be significantly improved [141, 189]. However, due to the constraint of Eq. 8, optimality-invariant augmentations cannot provide sufficient diversity to enhance the agent's generalization ability [55, 189]. Consequently, it is necessary to break the limitation of optimality invariance to capture the variation between the training and test environments [55].

2.3.2 Prior-Based Diversity

Based on prior knowledge about task-irrelevant contextual variations between training and test environments, targeted augmentations can be strategically applied to capture these variations effectively [83]. This approach introduces *prior-based diversity* by modifying corresponding features in the observed images. It is important to note that while DA can manipulate observed images, it cannot directly alter the distribution of the latent context. Figure 7 illustrates this concept using a representative example from DMControl-GB [54].

Knowing that background color and style vary between training and test environments, we can deliberately employ

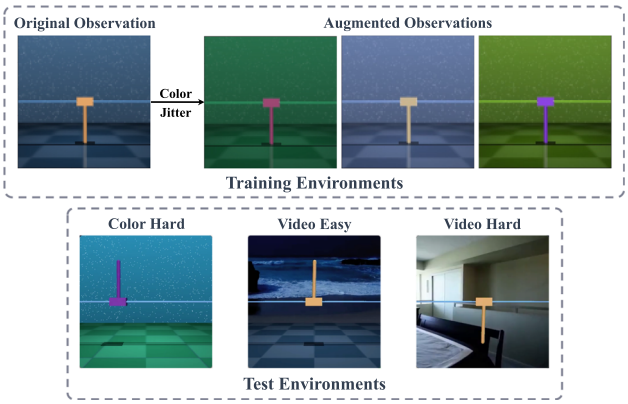


Fig. 7 Examples of applying DA under the assumption of prior-based diversity, where known task-irrelevant features such as background and color undergo substantial variations based on prior knowledge.

augmentation techniques such as color jitter to diversify training observations. Through this approach, agents can learn to identify task-irrelevant features by developing either an invariant policy or a robust latent representation from prior-based strong augmentation [83].

Strong augmentation under the prior-based diversity assumption breaks the limitation of the optimality invariance assumption and therefore has tremendous potential for improving the agent's generalization ability. However, this approach inevitably increases the estimation variance of the Q-values and thus may harm the stability of the RL optimization process [35, 55].

3 How to Augment Data in Visual RL?

The aim of DA is to increase the amount and diversity of the original training data so that agents can learn more efficient and robust policies [55]. Thus, a primary focus of previous research was to design effective augmentation approaches [168, 194]. In this section, we introduce the mainstream augmentation techniques and discuss the pros and cons of these methods.

Based on the type of data being augmented, we categorize the DA approaches in visual RL into three main types, as illustrated in Fig. 8. The first category, **observation augmentation**, involves transforming the given observations while keeping other transition factors (e.g., actions and rewards) unchanged, similar to label-preserving perturbations in SL. In Section 3.1, we detail various methods for employing DA on observations, which include not only diverse classical image manipulations directly applied to observation inputs but also several examples of DA in the feature space. The other two types, **transition augmentation** and **trajectory augmentation**, specifically take into account the unique properties of RL to broaden the scope of augmentation. In Section 3.2, we introduce transition augmentation, which enhances observations along with supervision signals, such as rewards. Finally, in Section 3.3, we explore trajectory augmentation, focusing on generating synthesized sequential trajectories.

In addition to summarizing techniques for augmenting different data types, this section will also introduce three advanced DA techniques that enhance the diversity of DA and improve its overall effectiveness. **Automatic augmentation** aims to automatically select the optimal DA type based on the specific task (Section 3.4), and **task-aware augmentation** (Section 3.5) focuses on providing data diversity while preserving critical information within the data. Furthermore, in light of recent advancements in generative AI, contemporary research has explored the use of technologies such as GANs and diffusion models for data generation; we will discuss this **generative augmentation** approach in Section 3.6.

3.1 Observation Augmentation

A typical observation augmentation approach is to apply the classical image manipulations to the observed images; most such manipulations were originally proposed for computer vision applications. Following the taxonomy of [150], we identify five categories of image manipulations: geometric transformations (Section 3.1.1), photometric transformations (Section 3.1.2), noise injections (Section 3.1.3), random erasing (Section 3.1.5) and image mixing (Section 3.1.4). Figure 9 shows a list of the visualized examples.

3.1.1 Geometric Transformations

Geometric transformations, which maintain optimality invariance or label-preservation [150], are commonly employed to address the limited availability of training data. **Random cropping** is an effective preprocessing technique for improving data efficiency; it works on image data with mixed width and height dimensions by locating a random central patch in each frame with a specific dimensionality [189, 190]. In many visual RL scenarios, such as robotic manipulation tasks, the vital regions are often positioned at the centers of the images, and cropping can remove irrelevant edge pixels to simplify the learning process [88]. Similar to cropping, the **window** transformation selects a random region and masks out the cropped part of the image, while **translation** renders the image with a larger frame and randomly moves the image within that frame.

Various other geometric transformations have been explored in visual RL scenarios. For instance, **rotation** transforms images by rotating them r degrees clockwise or counterclockwise, where r is randomly sampled from a predefined range [88]; **flipping** augments the dataset through horizontal or vertical reflection of observations. Although

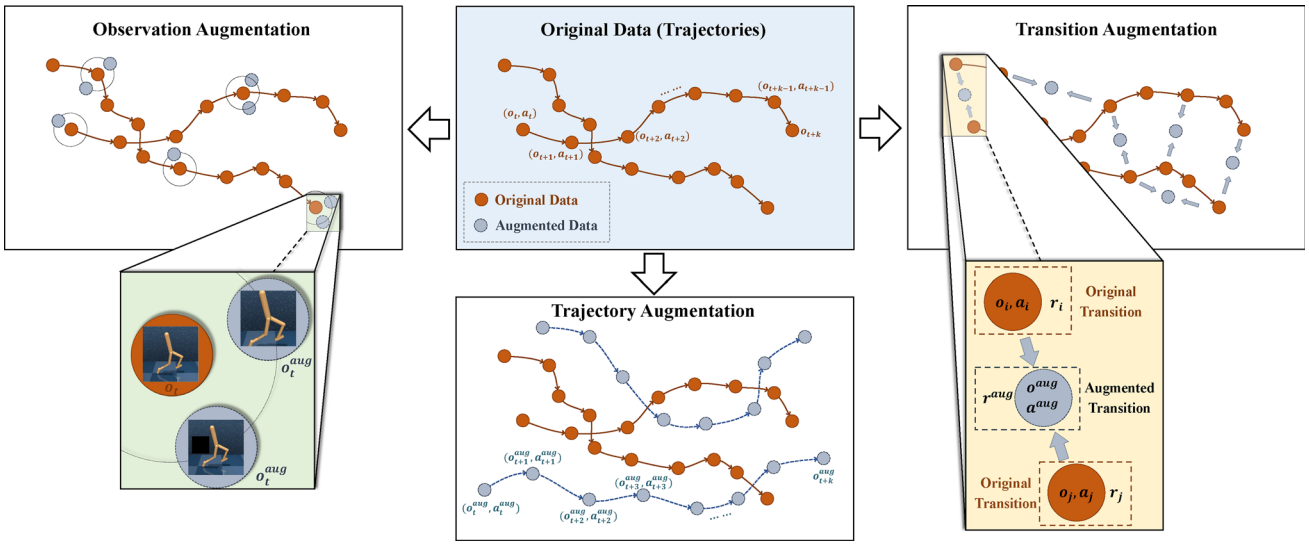


Fig. 8 The comparison of different DA paradigms depending on the type of data augmented: observation augmentation only generates synthetic observations o_t^{aug}; transition augmentation augments observations together with supervision signals $(o_t^{aug}, a_t^{aug}, r_t^{aug})$; and trajectory augmentation generates virtual trajectories $(o_t^{aug}, a_t^{aug}, o_{t+1}^{aug}, a_{t+1}^{aug}, \ldots, o_{t+k}^{aug})$.

proven effective in computer vision tasks, these transformations require careful consideration in visual RL as they may result in incorrect behavior without proper action adjustment to account for orientation changes.

3.1.2 Photometric Transformations

In real-world applications, object and background colors naturally vary due to environmental conditions such as lighting and weather [78]. Photometric transformations are designed to simulate these natural color variations, serving as a defense against overfitting to specific training data characteristics [126, 177]. This overfitting problem is particularly severe in visual RL, where agents may learn spurious correlations between task-irrelevant features and their policies, leading to significant performance degradation during testing [154]. To address this challenge, photometric transformations leverage prior knowledge about the variations between training and test environments to enhance the generalization of agent policies to unseen visual scenarios.

Several photometric transformation techniques have been developed: **grayscale** performs a straightforward RGB to grayscale conversion [88]; **color jitter** manipulates common image attributes including brightness, contrast, and saturation [25], typically implemented by converting images to HSV space and introducing controlled noise in the HSV channels [88]. Additionally, **random convolution** was introduced to address visual biases in convolutional neural networks (CNNs) [93]. This approach processes input observations through a randomly initialized single-layer convolutional network while maintaining input dimensions, effectively augmenting color information.

3.1.3 Noise Injection

Adding noise to images can help CNNs learn robust features in computer vision tasks [122], and recent studies [35, 55] also attempted to exploit this mechanism in visual RL to obtain robust state representations. In practice, distortion can be introduced by adding **Gaussian** noise [88] or **impulse** (salt-and-pepper) noise [35].

3.1.4 Image Mixing

This type of methods is commonly used in computer vision tasks to improve a model's robustness and generalization ability [209]. Among the different versions of mixing, **Overlay/Mixup** [207] trains a neural network on the convex combinations of samples and their labels. In visual RL, there are two ways to leverage the Mixup mechanism. First, we can combine two observations and their supervision signals, which will be discussed in Section 3.2. Alternatively, we can mix RL observations and other images randomly sampled from another dataset while the supervision signals of the observations remain fixed. For example, SECANT [35] linearly blends an observation with a distracting image I as $f(o) = \alpha o + (1-\alpha)I$, where I is randomly sampled from the COCO [101] image set.

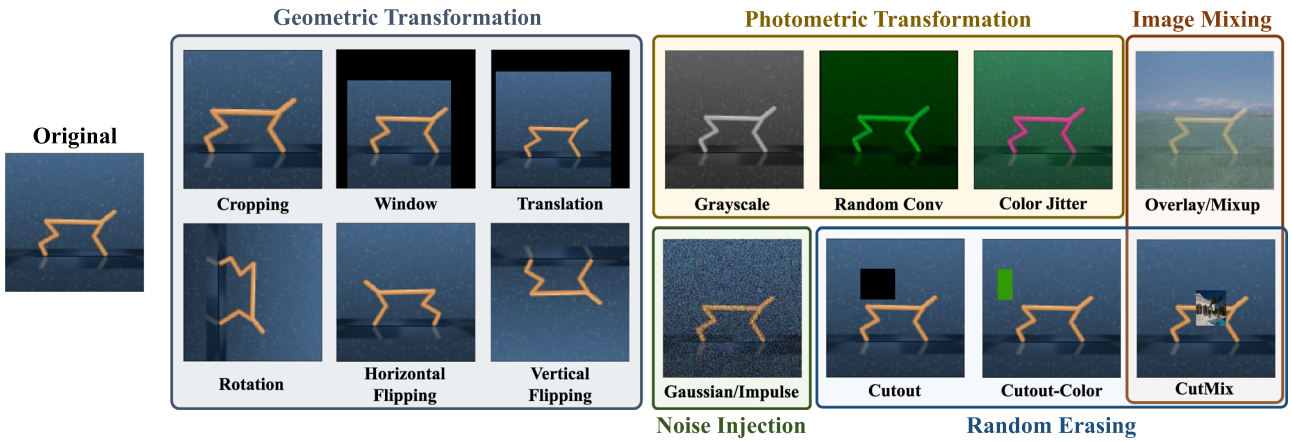


Fig. 9 Visualized examples of observation augmentation via classical image manipulations.

3.1.5 Random Erasing

Similar to dropout regularization, erasing techniques prevent overfitting by operating on input data rather than network architecture [212]. Several variants have been developed: **Cutout**[29] introduces random occlusions by masking an $m \times n$ patch of the input image; **Cutout-Color** extends this approach by filling the masked region with randomly sampled colors; **CutMix**[202], combining the principles of Cutout and Mixup, replaces the masked region with a patch from another image while maintaining the original supervision signals in visual RL settings [35].

3.1.6 Feature Space Augmentation

Beyond input-space transformations, an alternative approach is to perform augmentations in the feature space [28]. This feature space, also known as the latent space or embedding space, represents an abstract domain where meaningful representations of high-dimensional data are encoded.

Feature space augmentation primarily operates through two approaches. The first leverages autoencoders to map input images into latent features and reconstruct them after augmentation. Common techniques in this space include Gaussian noise injection and linear interpolation [21], which have demonstrated superior diversity compared to traditional transformations across various supervised learning tasks [106, 134]. While autoencoders have been employed in visual RL for reconstruction-based auxiliary tasks to enhance representation learning [127, 192], their potential for generating high-quality augmented data remains largely unexplored in this domain.

An alternative approach involves extracting and directly augmenting representations from the lower layers of CNNs without the need for high-dimensional image reconstruction [13, 150]. For example, MixStyle [213] implements style mixing at bottom layers to simulate diverse visual styles [70], achieving robust cross-domain generalization on benchmarks such as CoinRun [23]. Recently proposed CLOP [13] introduces a novel augmentation strategy that permutes pixel positions in feature maps at the deepest convolutional layer while preserving channel consistency. This approach leverages the abstract, high-level features encoded in deep layers to enhance generalization without requiring auxiliary representation learning tasks.

3.2 Transition Augmentation

As shown in Fig. 10, augmenting s_t with fixed supervision signals (e.g., the reward r_t and action a_t) can be regarded as a form of local perturbation of the corresponding transition, representing a key example of observation augmentation discussed in Section 3.1. To ensure the validity of the augmented transition $< s_t^{aug}, a_t, r_t, s_{t+1}^{aug} >$, the augmented observation s_t^{aug} must remain within a close range of the original observation s_t. As a result, local perturbation techniques face inherent limitations in expanding data diversity, a fundamental challenge shared across all observation augmentation approaches.

An intuitive solution is to apply interpolation across different data points instead of performing a local perturbation on each individual data point. Inspired by Mixup [207] and CutMix [202], MixReg [168] convexly combines two observations and their supervision signals to generate augmented data. For example, let y_i and y_j denote the signals for states s_i and s_j, respectively, which can be the reward or state values. After interpolating the observations by $\tilde{s} = \lambda s_i + (1-\lambda) s_j$, MixReg introduces mixture regularization in a similar manner via $\tilde{y} = \lambda y_i + (1-\lambda) y_j$, which helps learn more effective representations and smoother policies.

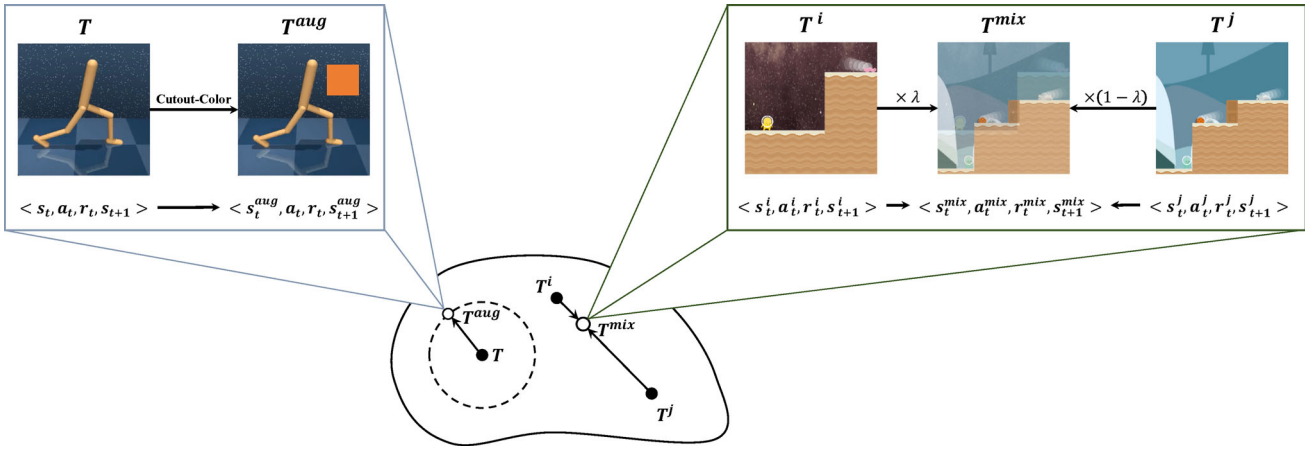


Fig. 10 Contrast between augmenting observations via local perturbations (left, Cutout-Color [29]) and augmenting observations with the supervision signals through interpolation (right, MixReg [168]).

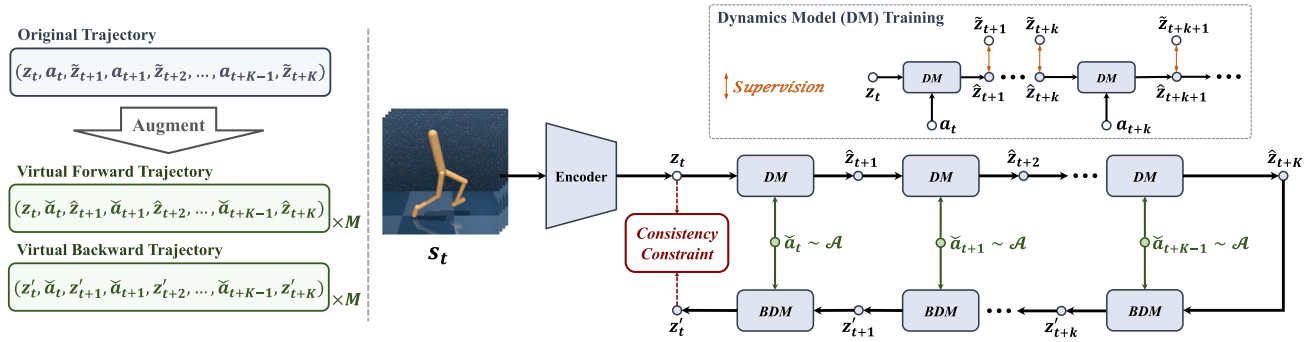


Fig. 11 Data flow and architecture of PlayVirtual [194] as an example of trajectory augmentation.

3.3 Trajectory Augmentation

Since observation or transition augmentation cannot directly enrich the trajectories encountered during training, to further improve the sample efficiency, PlayVirtual [194] augments the actions to generate synthesized trajectories under a self-supervised cycle consistency constraint.

In Fig. 11, PlayVirtual operates entirely in the latent space after encoding the input observation s_t into a low-dimensional state representation z_t. Following the dynamics model (DM) in SPR [144], PlayVirtual introduces a backward dynamics model (BDM) to predict the backward transition dynamics $(z_{t+1}, a_t) \longrightarrow z_t$ to build a loop with the forward trajectory. During the training process, the DM is supervised by the original trajectory information, whereas the BDM is constrained by the cycle consistency between z_t and z'_t. Further discussion on how to train the dynamics models with the auxiliary loss will be provided in Section 4.2. After obtaining the effective DM and BDM, PlayVirtual can generate diverse synthesized trajectories by randomly sampling/augmenting M sets of actions in the action space $\mathcal{A}$ and then calculating the state information. Experimental studies confirmed that regularizing feature representation learning with cycle-consistent synthesized trajectories is the key to PlayVirtual's success.

3.4 Automatic Augmentation

Automatic augmentation is receiving increasing attention due to the demand for task-specific augmentations [20, 25, 98]. For example, although random cropping is one of the most effective augmentation techniques for improving sample efficiency on many benchmarks, such as DMControl-500k [88, 189] and Procgen [141], the induced generalization ability improvement heavily depends on the specific choice of augmentation strategy. In general, different tasks benefit from different augmentations, and selecting the most appropriate DA method often requires expert knowledge. Consequently, it is crucial to develop methods that can automatically identify the most effective augmentation techniques. Research in visual RL remains in its early stages [141], and we highlight some promising approaches below.

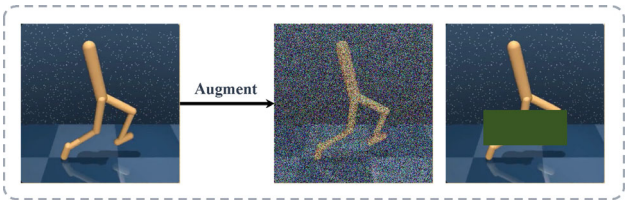


Fig. 12 Examples of context-agnostic augmentation, where augmentation operations inadvertently distort or eliminate critical control information.

1. **Upper Confidence Bound (UCB):** The task of selecting an appropriate augmentation from a given set can be formulated as a multi-armed bandit problem where the action space is the set of available transformations $F = \{f_1, f_2, \ldots, f_n\}$. The UCB [7] is a popular solution for the multi-armed bandit problem that considers both exploration and exploitation. Recently, UCB-DrAC [141] and UCB-RAD [42] were proposed to achieve automatic augmentation in visual RL. The experiment results suggest that UCB-based automatic augmentations can effectively improve the agent's generalization capabilities.
2. **Meta learning:** Meta learning offers an alternative solution to automatic augmentation and can be implemented in two ways [141]: (1) training a meta learner, such as RL^2[167], to automatically select an augmentation type before each update in a DA-based algorithm; (2) meta-learning the weights of a CNN to perturb observed images, a technique similar to model-agnostic meta learning (MAML)[16, 39]. In practice, neither approach has yielded promising results, and designing expressive functions for automatic augmentation via meta learning remains a challenge.

3.5 Context-Aware Augmentation

A notable limitation of existing DA techniques is their reliance on pixel-level image transformations that process each pixel without considering its contextual significance [198]. In the context of visual RL, however, pixels within an observation typically exhibit differential relevance to the decision-making process [45, 123]. As illustrated in Fig. 12, context-agnostic augmentation techniques may inadvertently mask or alter critical regions in the original observation that are essential for decision making.

This disregard for context elucidates why the straightforward application of prior-based strong augmentation, despite its potential to improve generalization, can significantly impair both sample efficiency and training stability in visual RL [55, 198]. Consequently, incorporating context awareness into augmentation techniques is essential for enhancing the effectiveness of DA while minimizing its potential drawbacks. Currently, two viable approaches have been proposed to advance context-aware augmentation:

1. **Introducing human guidance.** Human-in-the-loop RL (HIRL) [210] is a general paradigm that leverages human guidance to assist the RL process. EXPAND [47] introduces a human saliency map to mark the importance levels of different regions, and it only perturbs the irrelevant regions. Saliency maps contain human domain knowledge, allowing context information to be embedded into the augmentation.
2. **Excavating task relevance.** In visual RL, the contextual information can be extracted from the task relevance of each pixel, making it possible to directly determine its task relevance to achieve context-aware augmentation. Task-aware Lipschitz DA (TLDA) [198] explicitly defines the task relevance by computing the Lipschitz constants produced when perturbing corresponding pixels. Regions with large Lipschitz constants are crucial for the current task decision, and these regions will subsequently be protected from augmentation.

Context-aware augmentation forms the foundation for semantic-level DA, which aims to apply targeted operations to different semantic contexts within observations [44, 198]. In Section 6.1, we will further discuss semantic-level DA as a challenging yet pivotal direction for future research.

3.6 Generative Augmentation

Despite the remarkable success of leveraging generative models for data augmentation in computer vision tasks [5, 22, 116, 186, 211], the application of VAEs or GANs to generate synthetic data for reinforcement learning has not only failed to achieve comparable performance but may even lead to detrimental effects [74, 216]. This limitation in visual RL remained unresolved until the recent emergence of diffusion models. Several pioneering works have successfully demonstrated the potential of diffusion models in generating high-quality synthetic data for visual RL, marking a significant advancement in this domain.

1. **Generative Augmentation for Observations.** Recently, ROSIE [196] and GenAug [18] leverage text-guided diffusion models to augment observations in robotic control tasks while preserving the corresponding actions, representing an advanced approach to observation augmentation using generative models. Trained on massive online datasets, the diffusion models employed for DA can zero-shot create realistic images of many different objects and scenes. This approach enables significantly richer observation diversity compared to traditional DA techniques.

2. **Generative Augmentation for Transitions.** In contrast to approaches that solely generate observations, another line of research focuses on modeling the entire transition, simultaneously synthesizing novel action and corresponding reward labels. Within this paradigm, SynthER [109] directly trains diffusion models using either offline datasets or online replay buffers, subsequently generating samples for policy improvement. Advancing this concept further, MTDIFF [59] transcends the limitations of single-task scenarios by leveraging diffusion models to consolidate knowledge from multi-task datasets and augment data for novel tasks. The success of SynthER and MTDIFF demonstrates the significant potential of leveraging synthetic data to enhance visual RL performance.

Overall, recent studies and analyses indicate that data generated by diffusion models surpasses that of traditional DA methods in both diversity and accuracy [109]. This clearly demonstrates the capability of advanced generative models to produce novel, diverse, and dynamically accurate data. Such high-quality synthetic data can be effectively utilized by policies to enhance both the sample efficiency and generalization ability of visual RL algorithms. Furthermore, the text-controllable nature of current generative models enables them to serve as effective tools for semantically meaningful samples [18]. This capability holds promise for achieving genuine semantic-level manipulation of training data, presenting a crucial direction for future research.

3.7 Remarks

Data augmentation, as a data-centric approach, has demonstrated remarkable success in visual RL tasks, significantly enhancing both sample efficiency and generalization ability. This section provides a comprehensive review of various approaches addressing *"How to augment data in visual RL"*. Different augmentation strategies exhibit varying degrees of effectiveness across application scenarios. The optimal choice of augmentation technique often depends on the specific characteristics of the task at hand. Nevertheless, despite this task-specific nature, we can distill several fundamental insights that generalize across different contexts:

1. In contrast to DA in supervised learning scenarios, visual RL tasks encompass a richer set of data elements, including observations, latent states, and actions. Furthermore, these tasks incorporate temporal information that can be modeled as sequential transitions or complete trajectories, thus offering a wider range of manipulable data types [168, 194]. Among these, observation augmentation has gained the most widespread application due to its ease of implementation [190]. However, with the advancement of generative models, more complex yet diverse transition and trajectory augmentation techniques show potential for achieving novel breakthroughs.
2. To date, spatial perturbations and minor scaling of observations have emerged as the most effective and widely adopted DA approaches for enhancing sample efficiency in visual RL [113, 190]. This finding stands in marked contrast to conclusions from other domains, primarily attributable to the distinct underlying mechanisms through which DA enhances sample efficiency in visual RL [112]. We will discuss these mechanisms in depth in Section 6.3.
3. The robust generalization ability of visual RL agents during deployment largely depends on the diversity of training data. Consequently, both traditional strong augmentation techniques such as Color Jitter and advanced generative augmentation methods need to provide sufficient data richness to narrow the generalization gap [59]. In this context, ensuring training stability and maintaining the consistency of augmented data with environment dynamics become crucial considerations [55, 198].
4. The advancement of DA in visual RL hinges on developing methods for the automated generation of optimal, context-aware augmented data. Recent progress in has revealed promising avenues for leveraging pre-trained generative models to synthesize novel data that simultaneously exhibits rich diversity and maintains fidelity to the inherent constraints of RL dynamics.

4 How to Leverage Augmented Data in Visual RL?

Next, we discuss how to exploit the augmented data in visual RL. To ease the discussion, we divide the application scenarios where DA plays a vital role into three cases.

Case 1: **Sample-efficient RL in the single-environment setting.** Agents are trained and evaluated within a fixed environment, commonly referred to as the single-environment setting [190, 192]. The primary objective is to attain satisfactory performance with minimal interactions within the environment [113].

Case 2: **Generalizable RL in the multi-environment setting.** Agents are tested in unseen environments after interacting with the training environments [35, 96]. Since RL agents tend to overfit the training environment [206], generalizing the learned policies to unseen environments remains challenging even when only visual appearances are altered [3, 55].

Case 3: **Generalizable RL in the multi-task setting.** Agents in the multi-task setting aim to adapt to different tasks. Traditional end-to-end RL algorithms heavily rely on task-specific rewards, making them unsuitable for other

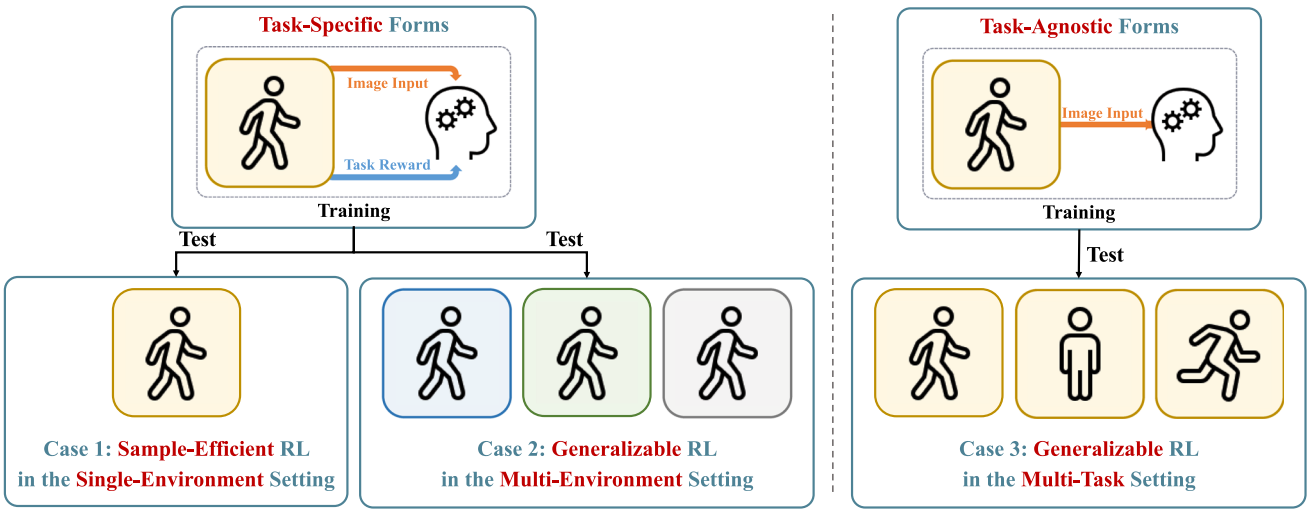


Fig. 13 Three representative scenarios highlighting the critical role of DA. In the context of single-task environments, DA enhances sample efficiency during training and improves generalization ability during deployment. Furthermore, DA contributes to training task-agnostic representations, facilitating superior generalization and adaptation across multiple tasks.

tasks [104, 156]. Recent studies have attempted to address this limitation by pretraining cross-task representations in a task-agnostic manner, thereby enabling agents to swiftly adapt to multiple downstream tasks [90, 200].

In Fig. 13, RL agents are trained with task-specific rewards in Case 1 and Case 2, where DA is implemented as an implicit regularization penalty when enlarging the training set (Section 4.1). However, the effect of implicit regularization is limited [144], and many studies have attempted to design auxiliary losses to exploit the potential of DA (Section 4.2). Some studies have also aimed to decouple representation learning from policy optimization to attain more generalizable policies [35] (Section 4.3). Finally, the related works belonging to Case 3, referred to as task-agnostic representation approaches using unsupervised learning, are introduced in Section 4.4.

4.1 Implicit Policy Regularization

DNNs are capable of learning complex representational spaces, which is essential for tackling intricate learning tasks. However, the model capacity required to capture such high-dimensional representations makes these techniques difficult to optimize and prone to overfitting [121]. Moreover, the complexity of visual RL is further aggravated by the need to jointly learn representations and policies directly from high-dimensional observations based on sparse reward signals [88, 192]. As a result, it is difficult for agents to distinguish the task-relevant (reward-relevant) features from high-dimensional observations, and they may mistakenly correlate rewards with spurious features [154]. To solve these issues, researchers have conducted a series of studies to develop effective regularization techniques, which can prevent overfitting and improve generalization by incorporating the inductive biases of model parameters [121].

In RL, a myriad of techniques have been proposed as regularizers such as L^p-norm regularization [100], batch normalization [36], weight decay [23] and dropout [72]. Among them, L^p-norm regularization explicitly includes regularization terms as additional constraints, and is referred to as explicit regularization [154]. Conversely, weight decay and dropout aim to tune the optimization process without affecting the loss function, making them implicit regularization strategies [72]. Additionally, DA has been prevalent in the deep learning community as a data-driven technique [65, 150]. Furthermore, increasing efforts have been devoted to the theoretical underpinnings behind DA [9, 14, 149, 184, 208] to explain its regularization effects, including the derivation of an explicit regularizer to simulate the behaviors of DA [9].

The initial and naive practice of DA is to expand the training set with synthesized samples [165]. This practice incorporates prior-based human knowledge into the data instead of designing explicit penalty terms or modifying the optimization procedure. Hence, it is often classified as a type of implicit regularization, formulated as the empirical risk minimization on augmented data (DA-ERM) [184] in SL tasks:

$$\widehat{h}^{da-erm} \triangleq \underset{h\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} l\left(h\left(\mathbf{x}_i\right), y_i\right) + \sum_{i=1}^{N}\sum_{j=1}^{\alpha} l\left(h\left(\mathbf{x}_{i,j}\right), y_i\right) \tag{9}$$

where $(\mathbf{x}i, y_i)$ represents the i^{th} original training sample ($\mathbf{x}i \in \mathcal{X}$ denotes the input feature, and $y_i \in \mathcal{Y}$ is its corresponding label); $\mathbf{x}i, j$ signifies the j^{th} augmented sample of $\mathbf{x}i$, which retains the corresponding label y_i; α indicates the number of augmentations; $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function, and $h(\cdot)$ is the model to be optimized.

In the visual RL community, RAD [88] and DrQ [189] first leverage classical image transformation strategies such as cropping to augment the input observations via the implicit regularization paradigm. In the original paper, DrQ is proposed with two distinct ways to regularize the Q-function. On the one hand, it uses K augmented observations from the original s_i' to obtain the target values for each transition tuple (s_i, a_i, r_i, s_i'):

$$y_i = r_i + \gamma \frac{1}{K} \sum_{k=1}^{K} Q_\theta(f(s_i', \nu_{i,k}'), a_{i,k}'),$$
$$a_{i,k}' \sim \pi(\cdot | f(s_i', \nu_{i,k}')) \tag{10}$$

where $f : \mathcal{S} \times \mathcal{T} \to \mathcal{S}$ is the augmentation function and ν is the parameter of $f(\cdot)$, which is randomly sampled from the set of all possible parameters $\mathcal{T}$. Alternatively, DrQ generates M different augmentations of s_i to estimate the Q-function:

$$J_Q^{\text{DrQ}}(\theta) = \frac{1}{NM} \sum_{i=1,m=1}^{N,M} ||Q_\theta(f(s_i, \nu_{i,m}), a_i) - y_i||_2^2 \tag{11}$$

In the above, DrQ leverages DA for improved estimation without adding any penalty terms, which is a type of data-driven implicit regularization. Since a sample can be defined as a tuple $(\mathbf{x}_i, y_i)$ in SL or a transition (s_i, a_i, r_i, s_i') in RL, the optimization objective of DrQ can be rewritten as:

$$J_Q^{\text{DrQ}}(\theta) = \frac{1}{NMK} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} l(f(s_i, \nu_{i,m}), a_i, r_i, f(s_i', \nu_{i,k}')) \tag{12}$$

where $l(s_i, a_i, r_i, s_i') = ||Q_\theta(s_i, a_i) - (r_i + \gamma Q_\theta(s_i', a_i'))||_2^2$ is the loss function, and $a_i' \sim \pi(\cdot|s_i')$. RAD [88] can be regarded as a specific form of DrQ with $K = 1$ and $M = 1$; it is a plug-and-play module that can be plugged into any RL method (on-policy methods such as PPO [143] and off-policy methods such as SAC [48]) without making any changes to the underlying algorithm. RAD has also highlighted the generalization benefits of DA [24].

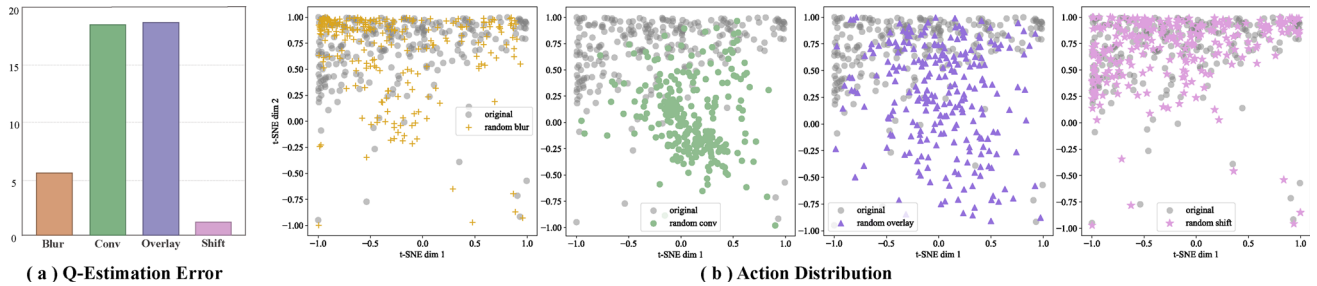
Since RAD and DrQ directly optimize the RL objective on multiple augmented observation views without any auxiliary losses, they can be viewed as implicit approaches for ensuring consistency and invariance among the augmented views. Building on DrQ, DrQ-v2 [190] makes several algorithmic adjustments, such as switching the baseline from SAC to DDPG and employing a larger replay buffer, which has resulted in significantly improved sample efficiency. The success of DrQ-v2 demonstrates that when utilizing weak augmentation to achieve sample-efficient visual RL algorithms, the method of implicit regularization can effectively harness the benefits of DA.

However, later studies found that implicit regularization with cropping exhibits poor generalization performance in unseen environments [55, 141]. As discussed in Section 2.3, optimality-invariant transformations (represented by cropping) cannot provide sufficient visual diversity for reducing the generalization gap. Furthermore, although prior-based strong augmentations such as color jitter have the potential to improve generalization, they may induce large Q-estimation errors and action distribution shifts, as shown in Fig. 14. Hence, implicit regularization approaches with prior-based strong augmentations (e.g., random convolution and overlay) may make the RL optimization process fragile and unstable [55, 198]. This poses a *dilemma* in visual RL: diverse augmentation is necessary to improve an agent's generalization ability, but excessive data variations may damage the stability of RL [35].

SVEA [55] aims to enhance the stability of RL optimization with DA [189]. It consists of two main components. First, SVEA uses only original data copies to estimate Q-targets to avoid erroneous bootstrapping caused by DA, where $y_i = r_i + \gamma Q_\theta(s_i', a_i'), a_i' \sim \pi(\cdot|s_i')$. Second, SVEA formulates a modified Q-objective to estimate the Q-value over both augmented and original copies of the observations, which can be expressed in a modified ERM form as follows:

$$\begin{aligned} J_Q^{\text{SVEA}}(\theta) =& \alpha \sum_{i=1}^{N} ||Q_\theta(s_i, a_i) - y_i||_2^2 \\ &+ \beta \sum_{i=1}^{N} \sum_{m=1}^{M} ||Q_\theta(f(s_i, \nu_{i,m}), a_i) - y_i||_2^2 \\ =& \alpha \sum_{i=1}^{N} l(s_i, a_i, r_i, s_i') \\ &+ \beta \sum_{i=1}^{N} \sum_{m=1}^{M} l(f(s_i, \nu_{i,m}), a_i, r_i, s_i') \end{aligned} \tag{13}$$

For actor-critic algorithms, SVEA employs strong augmentation exclusively during critic updates, with no augmentation applied during actor updates. SVEA assumes that the encoder's output embedding can become fully invariant to input augmentations. Under this assumption, an actor trained solely on unaugmented observations can indirectly



(a) Q-Estimation Error (b) Action Distribution

Fig. 14 Q-estimation errors and action distributions for augmented and original data. (a) Mean absolute Q-estimation errors of the converged DrQ [189] agents for the same observations before and after augmentation (copied from [55]). (b) Action distributions between the augmented and original data. We use t-distributed stochastic neighbor embedding (t-SNE) to show the high-dimensional actions employed by the same converged DrQ agent.

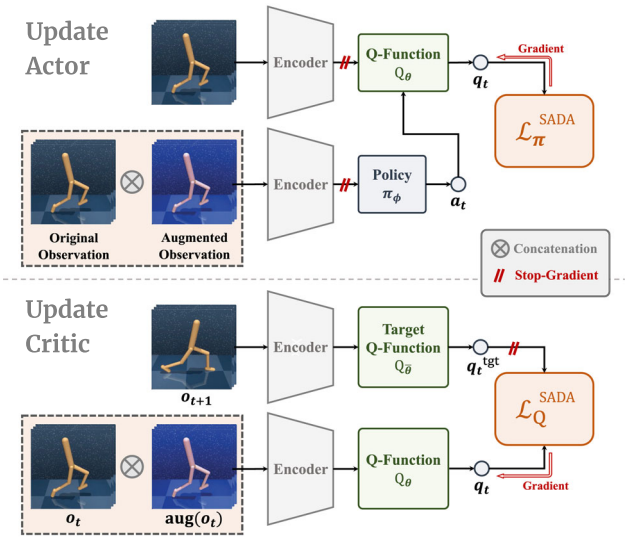


Fig. 15 The workflow of **S**tabilized Actor-Critic under Data Augmentation (SADA).

achieve robustness to augmented inputs via a shared actor-critic encoder. While this assumption holds for scenarios where the differences between test and training environments are limited to photometric changes, it fails when geometric augmentations are necessary for more complex generalization tasks.

To overcome this limitation, SADA [3] enhances the use of DA as implicit regularization to accommodate a broader range of augmentations. Instead of augmenting only the critic inputs, SADA carefully augments both actor and critic inputs to prevent training instabilities. As shown in the Fig. 15, (1) during actor updates, only the policy input is augmented while the Q-function input remains unchanged; (2) during critic updates, only the online Q-function input is augmented while the target Q-function input remains unaugmented; and (3) components are jointly optimized using both augmented and unaugmented data.

SVEA [55] and SADA [3] have significantly enhanced the generalization ability of visual RL algorithms. They achieve this by carefully designing pipelines that use strong augmentations without adding auxiliary losses or altering RL update mechanisms. This demonstrates the potential of implicit regularization to leverage DA for better generalization. However, it is crucial to meticulously design the use of strong augmentations to avoid training instabilities. In addition to carefully designing implicit regularization, other approaches have been explored to harness DA for improving both sample efficiency and generalization while avoiding adverse effects on training stability. These approaches include designing auxiliary tasks (Section 4.2) and decoupling representation learning from policy optimization (Section 4.3).

4.2 Explicit Policy Regularization with Auxiliary Tasks

Visual RL relies on the state representation, but it remains challenging to directly infer the ideal representation from high-dimensional observations [62]. A typical workflow involves designing auxiliary objectives to facilitate the representation learning process [11], or improve sample efficiency [89] or prevent observational overfitting [154]. In general, an auxiliary task can be considered an additional cost function that the RL agent predicts and observes from the environment in a self-supervised manner [140]. For example, the last layer of the network can be divided into multiple parts (heads), with each head dedicated to a specific task [97, 158, 193]. These multiple heads then propagate errors back to the shared network layers, thereby forming the comprehensive representations required by all heads.

With the recent success in unsupervised learning, various auxiliary tasks have been designed to produce effective representations [63, 140]. Thus, it is natural to design additional losses to explicitly constrain an agent's policy and value functions, which we will discuss in Section 4.2.1. Moreover, we introduce contrastive learning as a lower bound of mutual

information in Section 4.2.2 and future prediction objectives with a DM in Section 4.2.3.

4.2.1 DA Consistency

In contrast to simply inserting augmented data into the training dataset, DA consistency (DAC) [184] builds a regularization term to penalize the representation difference between the original sample $\phi_h(\mathbf{x}_i)$ and augmented sample $\phi_h(\mathbf{x}_{i,j})$, under the assumption that similar samples should be close in the representation space:

$$\underset{h\in\mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} l\left(h\left(\mathbf{x}_i\right), y_i\right) + \lambda \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{\alpha} \varrho\left(\phi_h\left(\mathbf{x}_i\right), \phi_h\left(\mathbf{x}_{i,j}\right)\right)}_{\text{DAC regularization}} \tag{14}$$

where ϕ_h refers to the features extracted from the high-dimensional data, which can be viewed as the output of any layer in the DNN, and ϱ is the metric function defined in the representation space, which can be the $\mathcal{L}_p$ norm or KL divergence. As an unsupervised representation module, DAC regularization can be employed as an auxiliary task in any SL or RL algorithms to enforce the model to produce similar predictions on the original and augmented samples. For example, SODA [54] calculates the consistency loss by minimizing the L^2 norm between the features of the augmented and original observations in the latent space; SIM [178] produces a cross-correlation matrix between two embedding vector sets of the original and augmented observations, and designs an invariance loss term to ensure the invariance of data.

For RL tasks, it is also desirable to train the network to output the same policies and values for both original and augmented observations [189]. For example, DrAC [141] employs two extra loss terms: G_π for regularizing the policy by the KL divergence measure and G_V for regularizing the value function using the mean-squared deviation:

$$\begin{aligned} G_\pi &= KL\left[\pi_\theta(a|s)|\pi_\theta(a|f(s,v))\right], \\ G_V &= \left(V_\phi(s) - V_\phi(f(s,v))\right)^2 \end{aligned} \tag{15}$$

The complete optimization objective of DrAC based on PPO is as follows:

$$J_{\text{DrAC}} = J_{\text{PPO}} - \alpha_r\left(G_\pi + G_V\right) \tag{16}$$

where α_r is the weight of the regularization term, and both G_π and G_V can be added to the objective of any actor-critic algorithm. By enforcing the DA consistency into the networks, specific transformations can be used to impose inductive biases relevant to the given task (e.g., invariance with respect to colors or translations) [141, 184].

Compared with implicit regularization techniques such as RAD and DrQ, DrAC employs two auxiliary consistency loss terms for explicitly regularizing the policy and the value function to ensure invariance. Instead of directly optimizing the RL objective on multiple augmented views of the observations, DAC regularization uses only the transformed observations $f(s,v)$ to compute the regularization losses G_π and G_V. Hence, DrAC can benefit from the regularizing effect of DA while mitigating the adverse effect on the RL objective [141].

4.2.2 Contrastive Learning

Another type of auxiliary task closely related to DA is contrastive learning. As **mutual information (MI)** is often hard to estimate, it is practical to maximize the lower bound of MI through approaches using, for example, InfoNCE loss [135]) to train robust feature extractors [164]. Recent studies [89, 156] have shown that contrastive learning can significantly improve the sample efficiency and generalization performance of visual RL [193]. Since contrastive learning only requires unlabeled data, it can not only be performed as auxiliary tasks together with RL objectives but also be leveraged to learn a task-agnostic representation, which we will discuss in Section 4.4.

In visual RL, there are two types of contrastive learning for improving agents' sample efficiency and generalization abilities [193]. The first class [34, 80, 89] focuses on **maximizing the MI between different augmented versions of the same observation** while minimizing the similarity between different observations. It tends to further exploit the regularization ability of DA at the MI level [193]. However, simply maximizing the lower bound of MI may retain the task-irrelevant information [37], which needs to be eliminated based on the information bottleneck principle. The second class [135, 156] aims to **maximize the predictive MI between consecutive states** by applying contrastive losses between an observation o_t and the near-future observations o_{t+k} over multiple time steps. This technique encourages the encoder to extract the temporal correlations of the latent dynamics from the observations [156], and DA can be applied as the prior-based data preprocessing.

Maximizing Multi-view MI: In self-supervised representation learning, feature extractors can be trained by maximizing the MI between different augmented views of the original data [164], and this approach has also been extended to the domain of visual RL [89, 157].

CURL [89] is the first general framework for combining multi-view contrastive learning and DA in visual RL. It builds an auxiliary contrastive task to learn useful state

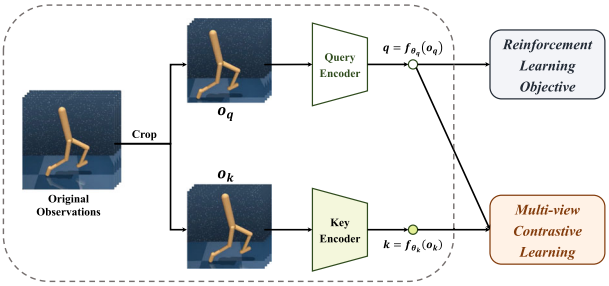


Fig. 16 The workflow of **c**ontrastive **u**nsupervised representations for **RL** (CURL).

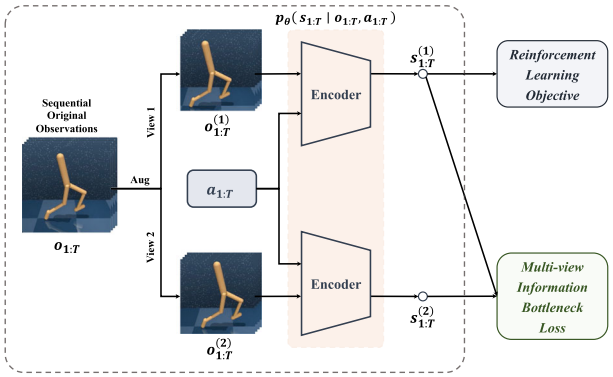


Fig. 17 Workflow of deep RL via multi-view **i**nfomration bottleneck (DRIBO).

representations by maximizing the MI between the different augmented views of the same observations to improve the transformation invariance of the learned embedding. In Fig. 16, the contrastive representation is jointly trained with the RL objective, and the latent encoder receives gradients from both the contrastive learning objective and the RL objective.

A key component of contrastive learning is the selection of positive and negative samples relative to an anchor, and CURL uses instance discrimination rather than patch discrimination [89]. Specifically, the anchor and positive observations are two different augmentations of the same observation, while the negative samples come from other observations in the minibatch. The contrastive learning task in CURL aims to maximize the MI between the anchor and the positives while minimizing the MI between the anchor and the negatives.

Following the setting of momentum contrast (MoCo) [60], CURL applies DA twice to generate queries and key observations, which are then encoded by the query encoder and key encoder, respectively. The query observations o_q are treated as the anchor, while the key observations o_k contain the positives and negatives. During the gradient update step, only the query encoder is updated, while the key encoder weights are set to the exponential moving average (EMA) of the query weights [60]. CURL employs the bilinear inner product $\text{sim}(q, k) = q^T W k$ to measure the agreement between query-key pairs, where W is a learned parameter matrix. Then, it uses the InfoNCE loss [135] to build an auxiliary loss function as follows:

$$\mathcal{L}_{\text{InfoNCE}} = \log \frac{\exp(q^T W k_+)}{\exp(q^T W k_+) + \sum_{i=0}^{K-1} \exp(q^T W k_i)} \tag{17}$$

where $\{k_0, k_0, \ldots, k_{K-1}\}$ are the keys of the dictionary and k_+ denotes a positive key. The InfoNCE loss can be interpreted as the log-loss of a K-way softmax classifier whose label is k_+ [164].

Many subsequently developed contrastive multi-view coding methods also employ the InfoNCE bound to maximize the MI between two embeddings that result from different augmentations. For example, DRIBO [34] aims to maximize the InfoNCE loss $\hat{I}_\psi(o_t^{(1)}, o_t^{(2)})$, where ψ represents the learnable parameters. Moreover, ADAT [80] selects the positive observations with the same action type and the negatives with other actions so that more positives can be produced. CCLF [157] introduces a curiosity appraisal module to select the most informative augmented observations for enhancing the effect of multi-view contrastive learning.

Although maximizing similarity between augmented observations benefits representation learning [89, 193], this approach of maximizing the mutual information lower-bound may inadvertently preserve task-irrelevant features, thereby limiting the agent's generalization capabilities.

To address this limitation, DRIBO [34] combines contrastive learning with a multi-view information bottleneck (MIB)-based auxiliary objective. The key insight is that effective RL representations should both facilitate future state prediction and minimize task-irrelevant information from visual observations. As illustrated in Fig. 17, DRIBO assumes that augmented observations share identical task-relevant information, while any unshared information is considered task-irrelevant [34].

In implementation, DRIBO maximizes task-relevant MI using InfoNCE (similar to CURL) while leveraging the information bottleneck principle to construct a relaxed Lagrangian loss. This approach aims to obtain sufficient representations with minimal task-irrelevant information. Specifically, the task-irrelevant minimization term is upper-bounded by:

$$\begin{aligned}\mathcal{L}_{\text{SKL}} = D_{\text{SKL}}(p\theta(\boldsymbol{st}^{(1)}|\boldsymbol{ot}^{(1)}, \boldsymbol{st}-1^{(1)}, \boldsymbol{at}-1) \\ ||p\theta(\boldsymbol{st}^{(2)}|\boldsymbol{ot}^{(2)}, \boldsymbol{st}-1^{(2)}, \boldsymbol{at}-1))\end{aligned} \tag{18}$$

where D_{SKL} denotes the symmetrized KL divergence computed from the encoder-generated probability densities of $\boldsymbol{st}^{(1)}$ and $\boldsymbol{st}^{(2)}$.

Empirical evaluations demonstrate that DRIBO achieves substantial improvements in both generalization and robustness on standard benchmarks including the DM control suite [161] and Procgen [24].

Maximizing Temporal Predictive MI: Predictive representation learning presents another powerful paradigm for learning effective representations that can be seamlessly integrated with DA. The first approach is to directly minimize the prediction error between the true future states and the predicted future states via a dynamic transition model, which will be discussed in Section 4.2.3. Another approach is to maximize the lower bound of the MI between the embeddings of consecutive time steps to induce predictive representations without relying on a generative model.

Early works such as CPC [135] and ST-DIM [4] employed temporal contrastive losses to maximize mutual information between previous and future state embeddings, though without incorporating DA. Recent approaches including ATC [156], CCFDM [129], and CoDy [193] advance this framework by applying DA to pre-encoded observations, thereby imposing inductive bias against task-irrelevant information.

CoDy [193], for instance, introduces $\mathcal{L}$TMI to maximize the InfoNCE bound on temporal mutual information between current state-action embeddings and subsequent state embeddings, aiming to enhance latent dynamics linearity. The implementation first randomly samples transition batches (st, a_t, s_{t+1}) from the replay buffer. These transitions are then encoded to obtain positive sample pairs (z_t^1, c_t, z_{t+1}). For each positive pair, negative samples are constructed by replacing z_{t+1} with features z_{t+1} from other pairs (z_t^1, c_t, z_{t+1}) within the same minibatch. Additionally, M-CURL [214] introduces a novel approach using bidirectional transformers to reconstruct masked observation features from contextual observations, capturing temporal dependencies through contrastive learning between reconstructed and original features.

4.2.3 Future Prediction with a DM

The motivation of future prediction tasks is to encourage state representations to be predictive of future states given the current state and future action sequence [105, 130, 193]. Instead of maximizing the MI between the current state and the future state using the InfoNCE loss [4, 135, 156], SPR [144] produces state representations by minimizing the prediction error between the **true future states** and the **predicted future states** using an explicit multi-step DM. As shown in Fig. 18, this approach also incorporates DA into the future prediction task, which enforces consistency across different views of each observation.

The DM $h(\cdot, \cdot)$ operates entirely in the latent space to predict the transition dynamics $(\mathbf{z}_t, a_t) \rightarrow \mathbf{z}_{t+1}$, where $\mathbf{z}_t = f(\mathbf{o}_t)$ is encoded by the feature encoder $f(\cdot)$ of the current input observation o_t. The prediction loss is computed by summing up the differences (errors) between the predicted representations $\hat{\mathbf{z}}_{t+1:t+K}$ and the observed representations $\tilde{\mathbf{z}}_{t+1:t+K}$:

$$\begin{aligned}\mathcal{L}_{\text{pred}} &= \sum_{k=1}^{K} d\left(\hat{\mathbf{z}}_{t+k}, \tilde{\mathbf{z}}_{t+k}\right) \\ &= -\sum_{k=1}^{K} \left(\frac{\tilde{\mathbf{z}}_{t+k}}{\|\tilde{\mathbf{z}}_{t+k}\|_2}\right)^{\top} \left(\frac{\hat{\mathbf{z}}_{t+k}}{\|\hat{\mathbf{z}}_{t+k}\|_2}\right)\end{aligned} \tag{19}$$

where the latent representation $\hat{\mathbf{z}}_{t+1:t+K}$ is computed *iteratively* as $\hat{\mathbf{z}}_{t+k+1} \triangleq h\left(\hat{\mathbf{z}}_{t+k}, a_{t+k}\right)$, starting from $\hat{\mathbf{z}}_t \triangleq \mathbf{z}_t \triangleq f_o(o_t)$, and $\tilde{\mathbf{z}}_{t+k} \triangleq f_m(o_{t+k})$ is computed by the target encoder f_m, whose parameters are the EMAs of the parameters of the online encoder f_o. Combined with DA, SPR improves the agent's sample efficiency and results in superior performance with limited iterations on Atari Games and the DeepMind control suite [144].

PlayVirtual [194] is an extension of SPR that introduces cycle consistency to generate augmented virtual trajectories for achieving enhanced data efficiency. Following the DM in SPR [144], PlayVirtual [194] proposes a BDM for backward state prediction to build a cycle/loop with a forward trajectory. Given a DM $h(\cdot, \cdot)$, a BDM $b(\cdot, \cdot)$, the current state representation $\mathbf{z}_t$, and a sequence of actions $a_{t:t+K}$, a forward trajectory and the corresponding backward trajectory can be generated to form a synthesized trajectory:

$$\begin{aligned}&\textit{Forward}: \hat{\mathbf{z}}_t = \mathbf{z}_t, \hat{\mathbf{z}}_{t+k+1} = h\left(\hat{\mathbf{z}}_{t+k}, \mathbf{a}_{t+k}\right), \\ &\qquad \text{for } k = 0, 1, \cdots, K-1 \\ &\textit{Backward}: \mathbf{z}'_{t+K} = \hat{\mathbf{z}}_{t+K}, \mathbf{z}'_{t+k} = b\left(\mathbf{z}'_{t+k+1}, \mathbf{a}_{t+k}\right), \\ &\qquad \text{for } k = K-1, K-2, \cdots, 0\end{aligned} \tag{20}$$

Since cycle consistency can be enforced by constraining the distance between the starting state $\mathbf{z}_t$ and the ending state $\mathbf{z}'_t$ in the loop, appropriate synthesized training trajectories can be obtained by augmenting actions. In practice, the cycle consistency loss can be calculated by randomly sampling M sets of actions from the action space $\mathcal{A}$:

$$\mathcal{L}_{\text{cyc}} = \frac{1}{M} \sum_{m=1}^{M} d_{\mathcal{M}}\left(\mathbf{z}'_t, \mathbf{z}_t\right) \tag{21}$$

where $d_{\mathcal{M}}$ is the distance metric over the latent space $\mathcal{M}$. The performance of PlayVirtual [194] can be explained from two aspects. First, the generated trajectories can help the agent "see" more flexible experiences. Second, enforcing the

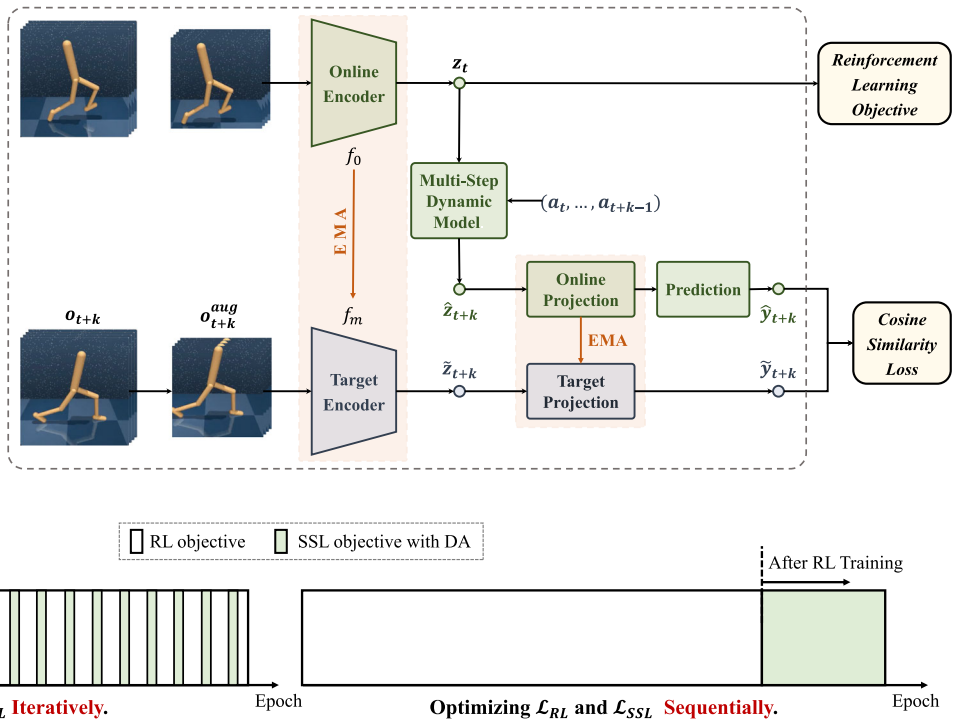


Fig. 18 The workflow of **s**elf-**p**redictive **r**epresentations (SPR).

Fig. 19 Different strategies for decoupling policy optimization and representation learning.

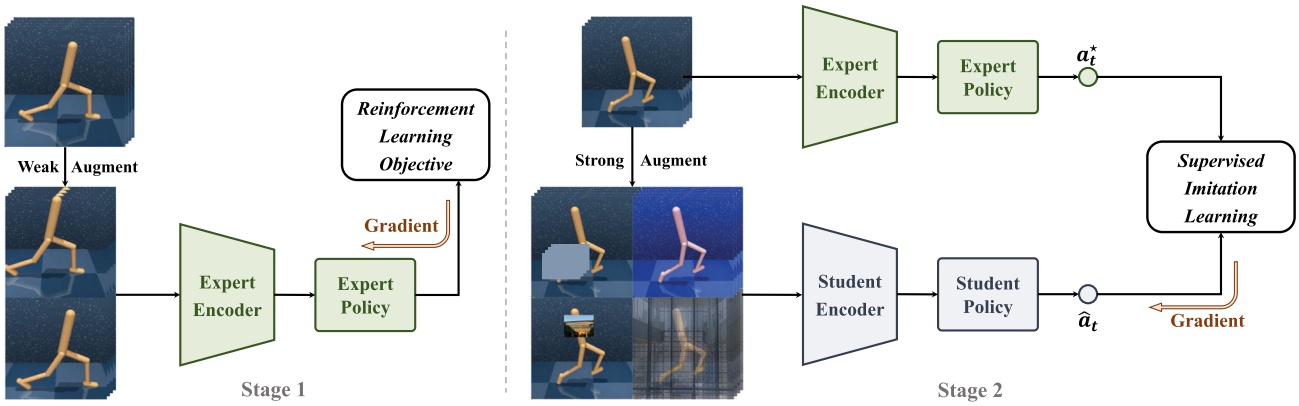


Fig. 20 The workflow of **s**elf **e**xpert **c**loning for **a**daptation to **n**ovel **t**est-env (SECANT).

trajectory with the cycle consistency constraint can further regularize the feature representation learning process.

4.3 Task-Specific Representation Learning Decoupled from Policy Optimization

Utilizing DA as an implicit [13, 88, 189] or explicit regularization approach with purposefully designed auxiliary tasks [89, 144, 193], the sample efficiency of visual RL has been significantly improved, resulting in performance comparable to state-based algorithms on several benchmarks [190]. However, training *generalizable RL agents* that are robust against irrelevant environmental variations remains a challenging task. Similar challenges in SL tasks, such as image classification, can be addressed by strong augmentations that heavily distort the input images, such as Mixup [207] and CutMix [202]. However, since the training process of RL is vulnerable to excessive data variations, a naive application of DA may severely damage the training stability [35, 55].

This poses a dilemma: aggressive augmentations are necessary for achieving good generalization in the visual domain [64], but injecting heavy DA into the optimization of an RL objective may cause deterioration in both the sample efficiency and the training stability [198]. Recent works [35, 54] argued that this is mainly due to the conflation of two

objectives: policy optimization and representation learning. Hence, an intuitive idea is to decouple the training data flow by using nonaugmented or weakly augmented data for RL optimization while using strongly augmented data for representation learning. As shown in Fig. 19, two strategies are available for achieving the decoupling goal: (1) dividing the training data into two streams to separately optimize $\mathcal{L}_{RL}$ and $\mathcal{L}_{SSL}$; and **iteratively** updating the model parameters by the two objectives [54]; (2) optimizing the RL objective $\mathcal{L}_{RL}$ first and then **sequentially** leveraging DA combined with SSL objective $\mathcal{L}_{SSL}$ for knowledge distillation [35].

Optimizing $\mathcal{L}_{RL}$ and $\mathcal{L}_{SSL}$ Iteratively: This strategy aims to divide the training data into two data streams and only uses the nonaugmented or weakly augmented data for the RL training process; it leverages strong augmentations under prior-based diversity assumptions to optimize the self-supervised representation objective and enhance the generalization ability of the model. In practice, this technique can be performed by iteratively optimizing the RL objective $\mathcal{L}_{RL}$ and the self-supervised representation objective $\mathcal{L}_{SSL}$ in combination with DA to update the network parameters. For example, SODA [54] maximizes the MI between the latent representations of augmented and nonaugmented data as the auxiliary objective $\mathcal{L}_{\text{SODA}}$, and ***continuously alternates*** between optimizing $\mathcal{L}_{RL}$ with nonaugmented data and $\mathcal{L}_{\text{SODA}}$ with augmented data. While a policy is learned only from nonaugmented data, SODA still substantially benefits from DA through representation learning [54].

Optimizing $\mathcal{L}_{RL}$ and $\mathcal{L}_{SSL}$ Sequentially: This is a two-stage training strategy, which first trains a sample-efficient agent using weak augmentations, and then enhances the state representation by auxiliary self-supervised learning or imitation learning with strong augmentations. For example, SECANT [35] first trains a sample-efficient expert with random cropping (weak augmentation). In the second stage, a student network learns a generalizable policy by mimicking the behavior of the expert at every time step but with a crucial difference: the expert produces the ground-truth actions from unmodified observations, while the student learns to predict the same actions from heavily corrupted observations, as shown in Fig. 20. The student optimizes the imitation objective by performing gradient descent on a supervised regression loss: $\mathcal{L}(o;\theta_s) = \|\pi_s(f(o)) - \pi_e(o)\|_F$, which has better training stability than the RL loss. Furthermore, conducting policy distillation through strong augmentations can greatly remedy overfitting so that robust representations can be acquired without sacrificing policy performance.

4.4 Task-Agnostic Representation Learning Using Unsupervised Learning

Unsupervised/self-supervised pretraining has demonstrated remarkable success across various domains by training models without explicit supervision [15, 61, 174]. This paradigm enables efficient downstream task adaptation through fine-tuning. Following this success, researchers have explored developing **unsupervised pretrained RL agents** capable of rapidly adapting to diverse test tasks in zero-shot or few-shot settings [104, 166].

Recent studies [156] have identified a critical limitation in standard end-to-end RL: visual representations learned through task-specific rewards often generalize poorly to other tasks. To address this, an alternative approach proposes **task-agnostic exploration for learning visual representations** without relying on task-specific rewards [156, 191]. This framework is particularly valuable in **multi-task settings** where different tasks, defined by distinct reward functions, share similar visual environments. For example, the *Walker* domain in the DeepMind control suite [161] encompasses various tasks including *standing*, *walking forward*, and *flipping backward*.

Two principal strategies have emerged for learning task-agnostic encoders that map high-dimensional inputs to compact representations. The first involves designing **unsupervised representation tasks**, as detailed in Section 4.2. The second approach focuses on **maximizing intrinsic rewards** derived from self-supervised objectives such as particle-based entropy and curiosity [56, 103, 104, 145, 191], encouraging meaningful behavioral patterns without external rewards. Leveraging DA's capacity for enhancing prior discrimination, many unsupervised pretraining approaches combine DA with auxiliary tasks to learn more effective representations. For instance, ATC [156] integrates random cropping with contrastive learning for task-agnostic representation learning, while APT [104] and SGI [145] employ DA in designing self-predictive tasks [115].

4.5 Remarks

As a data-centric method, DA is independent of specific RL baseline algorithms and can smoothly integrate with various techniques and training paradigms. When applied selectively to observations without altering other aspects of the algorithm, such as loss functions and training methodologies, DA acts as a form of implicit regularization. Conversely, incorporating auxiliary tasks to create a joint loss function while performing DA represents explicit regularization. In some auxiliary tasks, such as DA consistency regularization and multi-view contrastive learning, DA is an indispensable component. In other tasks, such as future prediction, DA serves as an enhancement. Furthermore, to mitigate training instabil-

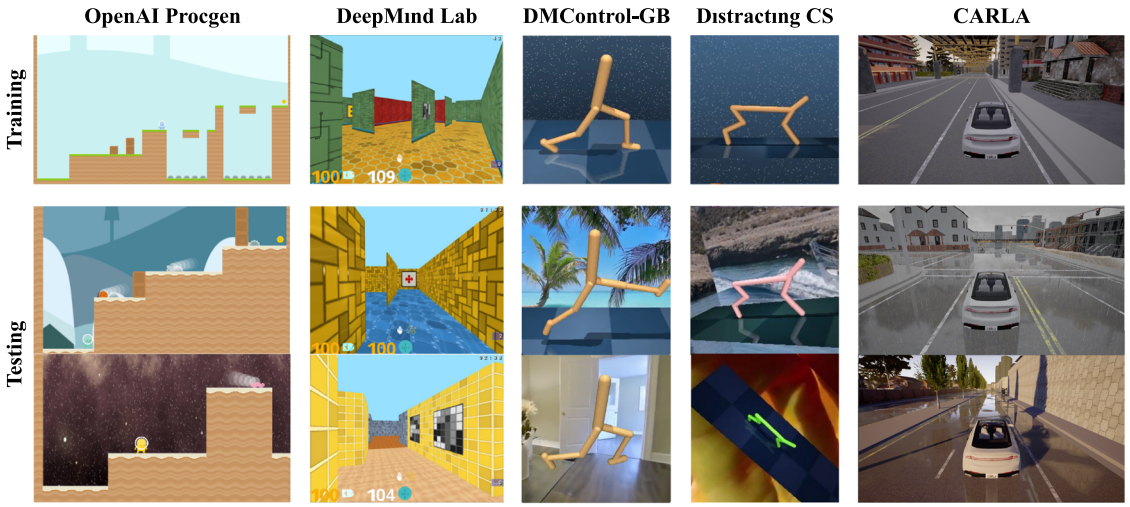


Fig. 21 Typical benchmarks used to evaluate an agent's generalization ability in visual RL.

ity induced by strong augmentation, the decoupling of visual representation learning from policy optimization is receiving growing attention. This approach proves advantageous for both task-specific representations and general representations that are not tied to specific tasks.

5 Experimental Evaluation

This section provides a systematic empirical evaluation of the methods in visual RL that leverage DA. First in Section 5.1, we introduce the commonly used benchmarks for evaluating the sample efficiency and generalization ability of agents. Then in Section 5.2 and Section 5.3, we present the experimental results of representative RL techniques using DA in comparison with those of other baselines to demonstrate the effectiveness of DA and identify the pros and cons of these methods.

5.1 Representative Benchmarks

5.1.1 Benchmarks for Sample Efficiency Test in Visual RL

Atari Games [75] This suite of games is widely used by both state-based and image-based discrete control algorithms for sample-constrained evaluations [120]. While RL algorithms can achieve superhuman performance on Atari games, they are still far less efficient than human learners, especially in image-based cases [89]. In the sample-efficient ***Atari-100k*** setting, only 100k interactions (400k frames with frame-skip=4) are available. The performance of an agent on a game is measured by its human-normalized score (HNS), defined as $\frac{S_A - S_R}{S_H - S_R}$, where S_A is the agent's score; S_R is the score of a random play; S_H is the expert human score.

DeepMind Control Suite [161] This is a continuous control benchmark suite for evaluating visual RL algorithms. It presents a variety of challenging tasks, including bipedal balancing, locomotion, contact forces, and goal reaching, with both sparse and dense reward signals. Previous studies usually measured the data efficiency and performance of their algorithms on the DeepMind control suites with 100k (for measuring learning speed) and 500k (for measuring overall performance) environment steps, which are referred to as ***DMControl-100k*** and ***DMControl-500k***, respectively [144]. DeepMind control suite is also a proper testbed for multi-task settings, as different tasks often involve the same domain. For example, the walker domain contains running, walking, standing and many other tasks, allowing agents to transfer learned policies to other tasks with similar visual observations.

5.1.2 Benchmarks for Generalization Test in Visual RL

Although Atari Games and the DeepMind control suite are suitable for benchmarking the sample efficiency of visual RL agents, they are not applicable for investigating the generalization abilities of these agents [88]. Generally, measuring the generalization ability of an agent requires variations between the training environment and the test environment, including state-space variations (the initial state distribution), dynamics variations (the transition function), visual variations (the observation function), and reward function variations [83]. In particular, DA-based techniques focus on zero-shot generalization to unseen environments with similar high-level goals and dynamics but different layouts and visual properties [13, 148]. Figure 21 shows the representative benchmarks for evaluating the agent's generalization ability in visual RL.

OpenAI Procgen [24] This is a suite of game-like environments where different levels feature varying visual attributes. Different combinations of the game levels can be used to separately construct training and test environments. Agents are only allowed to be trained on limited levels and are evaluated on unseen levels with different backgrounds or layouts [168].

DeepMind Lab [10] This is a first-person 3D maze environment in which various objects are placed in the rooms. As a measure of their generalization ability, agents are trained to collect objects in a fixed-map layout and tested in unseen environments that differ only in terms of their walls and floors (i.e., the variational contexts).

DeepMind Control Suite Variants [54, 83, 155] Since the original DeepMind control suite is not applicable for studying generalization, a number of variants have been proposed in recent years. Most of them, such as DMControl-GB [54], DMControl-Remastered [46] and Natural Environments [203], focus on visual generalization by changing the colors or styles of the background and floors. Furthermore, the Distracting Control Suite (DCS) [155] features a broader set of variations, including background style and camera pose variations. These variants provide meritorious benchmarks for evaluating the generalization abilities of continuous control algorithms using images as inputs.

CARLA [215] This is a realistic driving simulator where the agent's goal is to drive as far as possible in 1000 time steps without colliding into 20 other moving vehicles or barriers [35]. Learning directly from the rich observations in this scenario is challenging since diverse task-irrelevant distractors (e.g., lighting conditions, shadows, clouds, etc.) are available around the agent, which increases the difficulty of extracting control-related features.

5.2 Sample Efficiency Evaluation

To measure the sample efficiency of algorithms, we report the results on three common benchmarks: Atari-100k, DMControl-100k and DMControl-500k.

5.2.1 Atari-100k

In Table 1, the results of a random player (Random) and an expert human player (Human) are copied from [171] as baselines. Other scores are copied from their original papers [80, 144, 157, 189, 194, 214]. The results show that augmenting the observations as implicit regularization is effective, boosting the performance in terms of the median HNS from 5.8% (Efficient DQN) to 26.8% (DrQ). Moreover, appropriate auxiliary tasks such as contrastive learning [89, 214] and future prediction representation [144, 173, 194] can further yield improved sample efficiency. Among them, SPR [144] achieves the highest mean HNS value (70.4%) with its future prediction module, while PlayVirtual [194] achieves the highest median HNS value (47.2%) with the trajectory augmentation.

5.2.2 DMControl-100k and DMControl-500k

Compared with Atari games, the tasks in the DeepMind control suite are more complex and challenging. We first report the performance of the underlying SAC algorithm [48] based on state and image inputs, referred to as Pixel SAC and State SAC in Table 2 (copied from [89]), respectively, followed by the results of SAC-AE [192]. Since State SAC operates on low-dimensional state-based features instead of pixels, it approximates the upper bounds of sample efficiency in these environments for image-based agents. Similar to the case of Atari-100k, DrQ [189] achieves significant improvements over the underlying SAC algorithm [48], which is unable to complete these tasks. Combining auxiliary tasks with DA provides improved performance and potential for training sample-efficient agents. For example, based on SPR [144], recent studies have achieved superior performance by introducing cycle consistency constraints for more diverse trajectories (PlayVirtual [194]) or curiosity modules for better exploration (CCFDM [129]).

5.3 Zero-Shot Generalization Evaluation

In this subsection, we report the studies conducted on two benchmarks representing two different types of generalization: Procgen [24] for level generalization in arcade games, and DMControl-GB [54] for vision generalization in robot control tasks.

5.3.1 Level Generalization on Procgen

In Table 3, the results of RAD [88] and DrAC [141] are based on their most suitable augmentation types for different environments, and UCB-DrAC selects the most suitable type of DA as a multi-armed bandit problem. Based on the comparison of RAD [88] and its underlying PPO algorithm [143], it is evident that appropriate augmentations are beneficial in almost every environment. Additionally, explicitly regularizing the policy and value functions after performing augmentations (as in DrAC [141]) leads to further improvements. The outstanding results of CLOP [13] and DRIBO [34] highlight the remarkable potential of subtly designed representation learning methods to distinguish task-relevant information from task-irrelevant information.

Table 4 shows the best augmentation type for each game (copied from the original paper of DrAC [141]). Random cropping achieves the best performance on 9 out of 16 instances, while 5 out of 16 game environments benefit sig-

Table 1 Evaluation of Sample Efficiency on Atari-100k. We report the scores and the mean and median HNSs achieved by different methods on Atari-100k. The results are copied from the original works [80, 144, 157, 189, 194, 214].

Game	Human	Random	DQN	CURL	CCLF	ADAT	DrQ	M-CURL	SPR	PlayVirtual
Alien	7127.7	227.8	558.1	558.2	920.0	1029.7	771.2	**1151.6**	801.5	947.8
Amidar	1719.5	5.8	63.7	142.1	154.7	147.3	102.8	**182.2**	176.3	165.3
Assault	742.0	222.4	589.5	600.6	612.4	**749.4**	452.4	613.5	571.0	702.3
Asterix	8503.3	210.0	341.9	734.5	708.8	864	603.5	738.1	**977.8**	933.3
Bank Heist	753.1	14.2	74.0	131.6	36.0	164	168.9	220	**380.9**	245.9
Battle Zone	37187.5	2360.0	4760.8	14870.0	5775.0	21240	12954.0	**21600**	16651.0	13260.0
Boxing	12.1	0.1	−1.8	1.2	7.4	0.4	6.0	5.9	35.8	**38.3**
Breakout	30.5	1.7	7.3	4.9	2.7	4.5	16.1	5.7	17.1	**20.6**
Chopper Command	7387.8	811.0	624.4	1058.5	765.0	1106	780.3	**1138.9**	974.8	922.4
Crazy Climber	35829.4	10780.5	5430.6	12146.5	7845.0	21240	20516.5	20781.2	**42923.6**	23176.7
Demon Attack	1971.0	152.1	403.5	817.6	**1360.9**	851.9	1113.4	864.4	545.2	1131.7
Freeway	29.6	0.0	3.7	26.7	22.6	**29.7**	9.8	28.9	24.4	16.1
Frostbite	4334.7	65.2	202.9	1181.3	1401.0	1943.2	331.1	**2342.2**	1821.5	1984.7
Gopher	2412.5	257.6	320.8	669.3	**814.7**	601.2	636.3	453.8	715.2	684.3
Hero	30826.4	1027.0	2200.1	6279.3	6944.5	7259.2	3736.3	7360.6	7019.2	**8597.5**
Jamesbond	302.8	29.0	133.2	471.0	308.8	**635.7**	236.0	436.2	365.4	394.7
Kangaroo	3035.0	52.0	448.6	872.5	650.0	956.9	940.6	1691.4	**3276.4**	2384.7
Krull	2665.5	1598.0	2999.0	**4229.6**	3975.0	3502.9	4018.1	3240.9	3688.9	3880.7
Kung Fu Master	22736.3	258.5	2020.9	14307.8	12605.0	**19146**	9111.0	17645.6	13192.7	14259.0
Ms Pacman	6951.6	307.3	872.0	1465.5	1397.5	1075	960.5	**1758.9**	1313.2	1335.4
Pong	14.6	−20.7	−19.4	−16.5	−17.3	−15.1	−8.5	−8.9	−5.9	**−3.0**
Private Eye	69571.3	24.9	351.3	218.4	100.0	**388**	−13.6	321.6	124.0	93.9
Qbert	13455.0	163.9	627.5	1042.4	953.8	1578	854.4	1785	669.1	**3620.1**
Road Runner	7845.0	11.5	1491.9	5661.0	11730.0	12508	8895.1	12320	**14220.5**	13534.0
Seaquest	42054.7	68.4	240.1	384.5	550.5	251.6	301.2	481.1	**583.1**	527.7
Up N Down	11693.2	533.4	2901.7	2955.2	3376.3	3597.8	3180.8	4399.5	**28138.5**	10225.2
Mean HNS (%)	100	0	13.7	38.1	38.2	47.2	35.7	46.6	**70.4**	63.7
Median HNS (%)	100	0	5.8	17.5	18.1	20.6	26.8	34.0	41.5	**47.2**
# Superhuman	N/A	0	1	2	3	6	2	3	**7**	4
# SOTA	N/A	0	0	1	2	5	0	6	**7**	5

nificantly from photometric transformations, including color jitter and random convolution. For a detailed understanding of the connection between the properties of environments and augmentation types, Fig. 22 suggests that the visual differences between the training environment and the test environment act as a major factor when determining the best augmentation type. For example, the background styles of Climber vary significantly across different levels, and manipulating the color and other photometric factors is intuitively beneficial to generalization. By contrast, the different levels of the maze game Chaser share similar visual information but exhibit increasing difficulty. Consequently, applying photometric augmentations is likely to fail in this setting, which is consistent with the experimental results. In such cases, the appropriate augmentation type is usually random cropping, which is beneficial to sample efficiency and contributes to improving the generalization performance. In addition, CaveFlyer is uniquely friendly with the rotation augmentation, which is often destructive in other games. A closer check of the game shows that the major regions of the observations (except the gray areas) feature different positions and angles, and rotation can effectively narrow down the differences among them.

5.3.2 Vision Generalization on DMControl-GB

As a variant of the DeepMind control suite [161], DMControl-GB [54] aims to evaluate the generalization ability of an agent by changing the image color or replacing the background with another image set (Fig. 21). A comparison of the performance levels achieved in seen environments (Table 2) and unseen environments (Table 5) shows that

Table 2 Evaluation of Sample Efficiency on the DeepMind Control Suite. The reported scores (means and standard deviations) are achieved by different methods on DMControl-100k and DMControl-500k. The results are copied from their original works with 10 random seeds [89, 129, 144, 157, 189, 192–194].

DMControl 100k	Pixel SAC	SAC-AE	CURL	DrQ	SPR	CCLF	CoDy	MLR	CCFDM	Play Virtual	State SAC
Finger,	179	747	767	901	868	**944**	887	907	880	915	811
Spin	±166	±130	±56	±104	±143	**±42**	±39	±58	±142	±49	±46
Cartpole,	419	276	582	759	799	799	784	806	785	**816**	835
Swingup	±40	±38	±146	±92	±42	±61	±18	±48	±87	**±36**	±22
Reacher,	145	225	538	601	638	738	624	**866**	811	785	746
Easy	±30	±164	±233	±213	±269	±99	±42	**±103**	±220	±142	±25
Cheetah,	197	252	299	344	467	317	323	**482**	274	474	616
Run	±15	±173	±48	±67	±36	±38	±29	**±38**	±98	±50	±18
Walker,	42	395	403	612	398	648	**673**	643	634	460	891
Walk	±12	±58	±24	±164	±165	±110	**±94**	±114	±132	±173	±82
Ball in cup,	312	338	769	913	861	914	948	933	**962**	926	746
Catch	±63	±196	±43	±53	±233	±20	±6	±16	**±28**	±31	±91
500k											
Finger,	179	914	926	938	924	**974**	937	973	906	963	811
Spin	±166	±107	±45	±103	±132	**±6**	±41	±31	±152	±40	±46
Cartpole,	419	730	841	868	870	869	869	872	**975**	865	835
Swingup	±40	±152	±45	±10	±12	±9	±4	±5	**±38**	±11	±22
Reacher,	145	601	929	942	925	941	957	957	**973**	942	746
Easy	±30	±135	±44	±71	±79	±48	±16	±41	**±36**	±66	±25
Cheetah,	197	544	518	660	716	588	656	674	552	**719**	616
Run	±15	±50	±28	±96	±47	±22	±43	±37	±130	**±51**	±18
Walker,	42	858	902	921	916	936	**943**	939	929	928	891
Walk	±12	±82	±43	±45	±75	±23	**±17**	±10	±68	±30	±82
Ball in cup,	312	810	959	863	963	961	970	964	**979**	967	746
Catch	±63	±121	±27	±9	±8	±9	±4	±14	**±17**	±5	±91

although DrQ [189] is prominent in terms of sample efficiency, the diversity derived from the naive application of cropping is limited, and a significant generalization gap is induced when this approach is transferred to unseen environments. To provide sufficient visual diversity for generalization, it is necessary to use strong augmentations such as random convolution and overlay, as indicated by the results of follow-up studies [55].

SVEA [55] and TLDA [198] both significantly outperform DrQ [189] by focusing on stabilizing the training process when leveraging strong augmentation to optimize the representation and policy together. Another way to improve generalization is to decouple the unsupervised representation learning and the RL optimization process, either in an iterative manner (e.g., SODA [54] and SIM [178]) or in a sequential manner (e.g., SECANT [35]). Moreover, pre-trained encoders from off-the-shelf image datasets such as PIE-G [199] from ImageNet [27] also show attractive potential to provide generalizable representations in downstream tasks.

6 Discussion and Future Works

DA techniques have substantially improved the sample efficiency and generalization abilities of visual RL methods; however, many challenges remain to be addressed. In this section, we elaborate on these points and highlight key directions for future research, encompassing the opportunities, limitations, and underlying mechanisms of leveraging DA in visual RL.

6.1 Towards Semantic-Level DA

It can be considered as a kind of feature manipulation technique that alters the relative contributions of task-relevant and task-irrelevant features in the gradient update steps of

Table 3 Evaluation of Generalization Ability on Procgen. Agents are trained on the first 200 levels of each game and evaluated on unseen levels. The scores are copied from the original papers on UCB-DrAC [141] and DRIBO [34]. The mean and standard deviation values are calculated with 10 random seeds.

Game	PPO	RandFM	MixReg	RAD	DrAC	UCB-DrAC	CLOP	DRIBO
BigFish	4.0±1.2	0.6±0.8	7.1±1.6	9.9±1.7	8.7±1.4	9.7±1.0	**19.2±4.6**	10.9±1.6
StarPilot	24.7±3.4	8.8±0.7	32.4±1.5	33.4±5.1	29.5±5.4	30.2±2.8	**40.9±1.7**	36.5±3.0
FruitBot	26.7±0.8	24.5±0.7	27.3±0.8	27.3±1.8	28.2±0.8	28.3±0.9	29.8±0.3	**30.8±0.8**
BossFight	7.7±1.0	1.7±0.9	8.2±0.7	7.9±0.6	7.5±0.8	8.3±0.8	9.7±0.1	**12.0±0.5**
Ninja	5.9±0.7	6.1±0.8	6.8±0.5	6.9±0.8	7.0±0.4	6.9±0.6	5.8±0.4	**9.7±0.7**
Plunder	5.0±0.5	3.0±0.6	5.9±0.5	8.5±1.2	**9.5±1.0**	8.9±1.0	5.4±0.7	5.8±1.0
CaveFlyer	5.1±0.9	5.4±0.8	6.1±0.6	5.1±0.6	6.3±0.8	5.3±0.9	5.0±0.3	**7.5±1.0**
CoinRun	8.5±0.5	9.3±1.4	8.6±0.3	9.0±0.8	8.8±0.2	8.5±0.6	**9.6±0.1**	9.2±0.7
Jumper	5.8±0.5	5.3±0.6	6.0±0.3	6.5±0.6	6.6±0.4	6.4±0.6	5.6±0.2	**8.4±1.6**
Chaser	5.0±0.8	1.4±0.7	5.8±1.1	5.9±1.0	5.7±0.6	6.7±0.6	**8.7±0.2**	4.8±0.8
Climber	5.7±0.8	5.3±0.7	6.9±0.7	6.9±0.8	7.1±0.7	6.5±0.8	7.4±0.3	**8.1±1.6**
DodgeBall	**11.7±0.3**	0.5±0.4	1.7±0.4	2.8±0.7	4.3±0.8	4.7±0.7	7.2±1.2	3.8±0.9
Heist	2.4±0.5	2.4±0.6	2.6±0.4	4.1±1.0	4.0±0.8	4.0±0.7	4.5±0.2	**7.7±1.6**
Leaper	4.9±0.7	6.2±0.5	5.3±1.1	4.3±1.0	5.3±1.1	5.0±0.3	**9.2±0.2**	5.3±1.5
Maze	5.7±0.6	8.0±0.7	5.2±0.5	6.1±1.0	6.6±0.8	6.3±0.6	5.9±0.2	**8.5±1.6**
Miner	8.5±0.5	7.7±0.6	9.4±0.4	9.4±1.2	**9.8±0.6**	9.7±0.7	**9.8±0.3**	**9.8±0.9**

Table 4 Best augmentation types for DrAC [141] in different games. The original experiments [141] investigate a set of eight transformations: cropping, grayscale, Cutout, Cutout-Color, flipping, rotation, random convolution and color jitter (all of them are shown in Fig. 9).

Game	BigFish	StarPilot	FruitBot	BossFight	Ninja	Plunder	CaveFlyer	CoinRun
Aug Type	Crop	Crop	Crop	Flip	Color Jitter	Crop	**Rotate**	Random Conv
Game	Jumper	**Chaser**	**Climber**	DodgeBall	Heist	Leaper	Maze	Miner
Aug Type	Random Conv	**Crop**	**Color Jitter**	Crop	Crop	Crop	Crop	Color Jitter

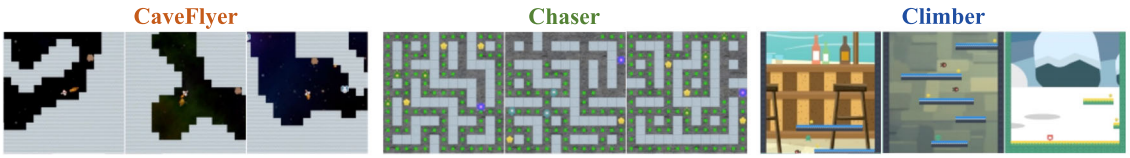


Fig. 22 Examples of three games with different structures. As shown in Table 4, the most effective augmentations are color jitter, rotation and cropping for Climber, CaveFlyer and Chaser, respectively.

the utilized network [149]. In this context, an ideal (albeit theoretical) DA method would operate at the semantic level, possessing the capability to precisely identify features pertinent to the current label or task while effectively perturbing irrelevant information.

However, this assumption underlying label-preserving transformations in SL and optimality-invariant transformations in RL proves challenging to uphold in practice, particularly when implementing pixel-level augmentations. The fundamental limitation of pixel-level approaches becomes evident when considering their mechanism: such augmentations, which aim to transform each pixel in a context-agnostic manner, inherently struggle to discriminate between task-relevant and task-irrelevant information [198]. This indiscriminate modification of pixels often results in the inadvertent alteration of critical task-relevant features, thereby compromising the efficacy of DA techniques, particularly in the domain of visual RL. Therefore, a promising avenue for advancing DA techniques lies in the development of semantic-level augmentation strategies. These strategies offer a more sophisticated approach compared to conventional pixel-level manipulations.

Several recent studies have attempted to move towards semantic-level DA by focusing on preserving task-relevant information. For example, EXPAND [47] and TLDA [198], as discussed in Section 3.5, propose methods to prevent

Table 5 Evaluation of Generalization Ability on DMControl-GB. The scores (means and standard deviations) are obtained by conducting training in a fixed environment and performing evaluations in two unseen test environments with random colors (top) and natural video backgrounds (bottom).

Random Colors	SAC	CURL	RAD	PAD	DrQ	SVEA	SODA	SIM	TLDA	SECANT	PIE-G
Walker,	365	662	644	797	770	942	930	940	**947**	939	941
Stand	±79	±54	±88	±46	±71	±26	±12	±2	**±26**	±7	±35
Walker,	144	445	400	468	520	760	697	803	823	856	**884**
Walk	±19	±99	±61	±47	±91	±145	±66	±33	±58	±31	**±20**
Cartpole,	248	454	590	837	630	837	831	841	760	**866**	749
Swingup	±24	±110	±53	±63	±52	±23	±21	±13	±60	**±15**	±46
Ball in cup,	151	231	541	563	365	**961**	949	953	932	958	**964**
Catch	±36	±92	±29	±50	±210	**±7**	±19	±7	±32	±7	**±7**
Finger,	504	691	667	803	776	**977**	793	960	–	910	–
Spin	±114	±12	±154	±72	±134	**±5**	±128	±6		±115	
Cheetah,	133	–	–	159	100	273	294	–	371	**582**	364
Run	±26			±28	±27	±23	±34		±51	**±64**	±40
Natural Videos											
Walker,	274	852	745	935	873	961	955	963	**973**	932	957
Stand	±39	±75	±146	±20	±83	±8	±13	±5	**±6**	±15	±12
Walker,	104	556	606	717	682	819	768	861	**873**	842	**870**
Walk	±14	±133	±63	±79	±89	±71	±38	±33	**±34**	±47	**±22**
Cartpole,	204	404	373	521	485	**782**	758	770	671	752	597
Swingup	±20	±67	±72	±76	±105	**±27**	±62	±13	±57	±38	±61
Ball in cup,	172	316	481	436	318	871	875	820	887	903	**922**
Catch	±46	±119	±26	±55	±157	±106	±56	±135	±58	±49	**±20**
Finger,	276	502	400	691	533	808	695	81	–	**861**	837
Spin	±81	±19	±64	±80	±119	±33	±94	±38		**±102**	±107
Cheetah,	80	–	–	206	102	292	229	–	356	**428**	287
Run	±19			±34	±30	±32	±29		±52	**±70**	±20

augmentation of salient or sensitive areas in observed images, thereby maintaining critical visual information. In the broader computer vision community, KeepAugment [44] uses a saliency map to identify the key regions and then preserves these informative regions during augmentation to produce reliable training samples. While not directly a DA technique, VAI [170] employs unsupervised keypoint detection and visual attention mechanisms, combined with a reconstruction loss, to compel the encoder to embed only the foreground information of the input image. This method effectively introduces an inductive bias based on the assumption that key information controlling the objective in observations typically resides in the foreground, thereby prioritizing task-relevant features.

Recent advancements in pre-trained generative models have opened new avenues for achieving semantic-level augmentation [18]. The robust generative capabilities of large multi-modal models offer a promising DA approach that can be controlled through prompts. For instance, these models can be instructed to "replace the background of a robotic arm with an outdoor scene while keeping the arm itself unchanged". Such prompt-driven DA transformations have the potential to generate diverse augmented data while retaining semantic invariance. The application of generative augmentation in visual RL is still in its nascent stages. This emerging field presents a fertile ground for further exploration, particularly in developing methods that can leverage the semantic understanding of these models to produce task-relevant augmentations. Future research directions should focus on three key aspects: developing RL-specific prompt engineering methodologies, systematically analyzing the impact of generated data on policy learning, and establishing mechanisms to ensure generated augmentations remain consistent with the underlying RL dynamics.

6.2 Trade-off between Training Stability and Generalization

In practice, DrQ [189, 190] and RAD [88] that leverage weak DA such as random cropping as implicit regularization

methods can yield significantly improved sample efficiency during training, while a noticeable generalization gap is observed when these approaches are transferred to unseen environments [35, 54, 55]. Furthermore, more diverse augmentations, such as color jitter, have the potential to improve generalization but tend to result in unstable optimization and poor sample efficiency [55, 88]. Therefore, a dilemma of balancing between stability and generalization is persistent when applying DA in visual RL. This challenge is particularly acute due to the inherently fragile nature of the optimization process in RL [35].

This dilemma is frequently attributed to the conflation of policy optimization and representation learning in current end-to-end visual RL algorithms [35, 54]. Consequently, a logical approach is to decouple these processes, independently learning a robust representation and a competent policy, as elaborated in Section 4.3 and Section 4.4. Such decoupling facilitates the application of heavy augmentations to improve generalization while simultaneously employing weak augmentations to maintain satisfactory sample efficiency [35, 54]. This strategy effectively addresses the stability-generalization trade-off. Moreover, the dilemma is exacerbated by the limitations of pixel-level augmentation techniques. These methods, when applied intensively, risk inadvertently destroying critical features, further complicating the balance between effective augmentation and preservation of essential information.

Further insight into this dilemma can be gained through the lens of the bias-variance trade-off, a fundamental principle in machine learning [30]. Contemporary complex models, such as DNNs, typically exhibit low bias but high variance. Consequently, these models are prone to overfitting the training data, resulting in sub-optimal performance on unseen data. DA addresses this issue by introducing increased diversity, thereby reducing variance and enhancing the model's generalization capabilities [206]. Although DA may mitigate the issue of overfitting, certain augmentation combinations can actually lead to underfitting, making the training process unstable and challenging [14, 121]. Note that the issue of underfitting is more detrimental in RL, as its optimization process is more unstable than those of supervised tasks.

Overall, there are two potential paths to further balance training stability and generalization. Firstly, designing more effective augmented data generation methods to avoid corrupting task-relevant information, as discussed in Section 6.1. Secondly, optimizing existing paradigms for leveraging DA, such as decoupling policy optimization and representation learning.

6.3 Underlying Mechanisms of DA Efficacy in Visual RL

Given the widespread application and powerful impact of DA in visual RL, understanding its underlying mechanisms has become crucial. However, most works have simply employed DA as a basic component without delving deeper into 'why DA works' [88, 189, 190]. To gain a comprehensive understanding of DA's effectiveness in visual RL, this section begins by briefly reviewing existing studies on the underlying mechanisms of DA within the broader context of deep learning (Section 6.3.1). Subsequently, we will explore in depth the unique mechanisms of DA's efficacy in visual RL tasks in Section 6.3.2.

6.3.1 Prevailing Perspectives on DA Effectiveness in DL

In recent years, numerous efforts have been made to investigate the theoretical guarantees for DA from various perspectives. These guarantees provide researchers with valuable insights into the practical effects of such approaches [8, 14, 149]. This section offers a concise overview of previous works on the theoretical foundations of DA, categorizing them into three main perspectives: implicit regularization [8, 65, 121], invariance learning [14, 119] and feature manipulation [149, 176].

Implicit Regularization vs. Explicit Regularization. Regularization is a fundamental technique in deep learning that aims to prevent overfitting and improve generalization abilities by constraining the complexity of a model [13, 121, 172]. The regularization strategies of DA act on the training data instead of the model's parameters and hence can be considered a type of implicit regularization approach instead of an explicit regularization technique that imposes constraints on the parameters, such as minimizing the $\mathcal{L}_2$ norms of the parameters [8]. By keeping the parameter space intact, this data-driven regularization approach can maintain the model's representational capacity while increasing its robustness [65, 189]. In fact, DA is more straightforward than explicit regularization that integrates the prior knowledge into objective functions, and neural networks can implicitly encode the attributes of DA without explicitly training towards these objectives [26, 180, 182]. Furthermore, attempts have also been made to derive explicit regularizers to describe the implicit regularization effect of DA [9, 67].

Transformation Invariance. Invariance is an essential property of all intelligent systems that makes them generalize effectively [12]. The purpose of DA is to constrain a model's output to be invariant when applying task-irrelevant transformations to the input data [141, 207]. It has been widely accepted that translation invariance is an inherent feature of CNNs [182], whereas other types of transformation

invariance, such as rotation invariance, must be induced by corresponding augmentations [73]. The definition of DA assumes that semantics are invariant to data transformation [150, 189], which implies that performing optimization with augmentation can result in implicit invariances. Furthermore, the specific details of augmentations can be used to encode prior knowledge about task-specific or dataset-specific invariances [12].

Feature Manipulation. An alternative explanation of how DA works is derived from the perspective of feature manipulation [187, 188, 217]. Learning meaningful features from high-dimensional data is empirically challenging, as critical features are often highly sparse and associated with spurious features such as dense noise. In practice, this may result in the network's overfitting the noisy features instead of properly learning the critical features [133]. First, by adjusting the relative contributions of the original data features, DA can effectively facilitate the incorporation of informative but hard-to-learn features into the learning process [149]. Second, the latest research shows that leveraging DA in contrastive learning can decouple spurious features from the representations of positive samples. By ignoring the decoupled features, the performance of networks may be boosted by focusing on the learning of resistant features [176].

6.3.2 Specialized Mechanisms of DA in Visual RL

The underlying mechanisms of DA in visual RL partially align with those observed in other domains, as discussed in Section 6.3.1. For instance, applying random shift to input observations without modifying other algorithmic details can be viewed as a form of implicit regularization [67, 113]. Additionally, the improvements in visual generalization achieved through DA can be attributed to the previously mentioned concepts of transform invariance and feature manipulation [83]. However, several remarkable phenomena unique to applying DA in visual RL cannot be adequately explained by mechanisms from other domains.

As illustrated by the blue training curve in Fig. 23, visual RL agents (based on the DDPG algorithm [100] in this example) fail to achieve an effective decision policy in the classic continuous control task *Walker Run* from the DeepMind Control suite [161]. Remarkably, merely applying simple random shift transformations to input observations, without any other modifications to the algorithm, results in superior performance [190], as demonstrated by the orange curve in Fig. 23. This striking improvement in training efficiency for visual RL stands in stark contrast to the incremental performance gains typically observed when applying DA in other task domains [185]. Such a significant disparity suggests that previous understandings of DA's mechanisms are insufficient to fully explain its role in visual RL tasks. Instead, this remarkable enhancement implies that DA must effectively overcome some critical bottleneck unique to visual RL that previously limited its training efficacy. This phenomenon motivates a deeper investigation into the underlying mechanisms through which DA contributes to the effectiveness of visual RL algorithms.

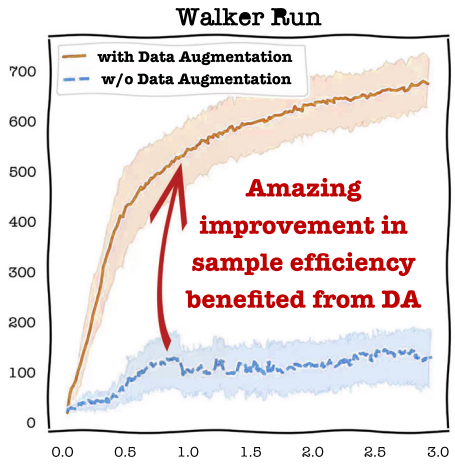


Fig. 23 Unlike in tasks from other domains, DA decisively enhances the training efficiency of visual RL.

The most recent investigation [112] has unveiled that DA's remarkable effectiveness in visual RL stems from its ability to mitigate the **plasticity loss** of deep neural networks. Plasticity, referring to the capacity of deep neural networks to continually learn from new data, gradually diminishes during training with non-stationary objectives [131, 153]. The inherent nature of DRL necessitates that agents continuously refine their policies through environmental interactions, resulting in intrinsically non-stationary data streams and optimization objectives [86]. This characteristic of DRL paradigms renders plasticity loss a critical impediment to achieving sample-efficient applications, as the networks must maintain their adaptability throughout the learning process [110, 132]. Compared to traditional state-based RL tasks, visual RL algorithms suffer from more severe plasticity loss due to increased task complexity and larger network architectures. It has been demonstrated through ingenious experiments [112] that without DA, agents fail to train effectively due to catastrophic plasticity loss, while applying DA significantly alleviates this bottleneck. Furthermore, DA's efficacy in mitigating plasticity loss surpasses that of several interventions specifically designed for this purpose, such as Layer Normalization [111], Shrink & Perturb [6], and CReLU [1].

6.4 Unique Characteristics of DA in Visual RL versus Other Domains

Given the widespread application of DA across various domains in deep learning [85, 207], a pertinent question naturally arises:

Does the implementation of DA in visual RL exhibit significant distinctions from its utilization in other domains, particularly in supervised and unsupervised vision tasks?

Indeed, there are substantial differences, which underscores the necessity of organizing a survey specifically focused on DA in visual RL. This section will delineate the most salient differences in DA implementation between visual RL and other domains. Further exploration of these distinctions and the development of tailored DA techniques for the visual RL scenario represent crucial directions for future research.

Diverse Augmentable Data Types. Compared to typical supervised or unsupervised vision tasks, RL data is inherently more complex. On one hand, RL data encompasses three distinct elements: state, action, and reward. On the other hand, RL involves long sequences of sequential decision-making data, known as trajectories. This complexity in data types enables a more diverse and flexible approach to transforming different data components in visual RL tasks. As introduced in Section 3, when categorizing based on the type of data being augmented, visual RL incorporates at least three classes of augmentation: observation augmentation, transition augmentation, and trajectory augmentation. Although existing works primarily utilize basic observation augmentation [88, 190], recent progress in generative models, particularly diffusion techniques, is poised to facilitate the development of more sophisticated and diverse DA strategies in visual RL [59, 216]. This paradigm shift toward sophisticated augmentation strategies shows great potential in exploiting the inherent structure of RL data more comprehensively, promising significant advances in both sample efficiency and generalization capabilities.

Distinctive Implementation Details. Due to the nonstationary nature of RL, there are significant differences in the optimal practices for applying DA in visual RL compared to other scenarios. Firstly, contrary to supervised vision tasks such as image classification, where heavy transformations like Mixup [207] and CutMix [202] demonstrate notable advantages over traditional image transformations, in visual RL, random cropping has emerged as the most practical augmentation technique [88, 141]. Consequently, the approach to harnessing augmented data must be thoughtfully designed to avoid potential destabilization of the optimization process while effectively exploiting the generalization capabilities induced by DA [113, 198]. Secondly, the timing of augmentation application is critical in RL [84], in contrast to supervised learning tasks [43], due to the heightened time sensitivity of augmentation in RL contexts. For example, optimality-invariant augmentations such as cropping should be implemented as early as possible to enhance sample efficiency and expedite the RL training process. Conversely, strong augmentations based on prior knowledge, exemplified by color jitter, may interfere with RL training stability, suggesting their optimal deployment during post-training knowledge distillation phases.

Unique Underlying Mechanisms. Building upon the discussion in Section 6.3.2, contrary to the general understanding of DA's mechanisms, it has been discovered that DA effectively addresses a bottleneck unique to RL: plasticity loss [112]. This insight explains the remarkable extent to which DA enhances the sample efficiency of visual RL algorithms [190]. The distinctive mechanism through which DA operates in this context necessitates a paradigm shift in our approach to designing DA methods for visual RL. Specifically, future work must adopt a novel perspective in approaching DA for visual RL by focusing on mitigating plasticity loss, a consideration that has been largely overlooked in traditional DA approaches. This conceptual shift represents a fundamental departure from conventional DA strategies, opening promising new research directions in visual RL.

6.5 The Role of Visual RL and DA in the Age of Foundation Models

In recent years, foundation models, particularly large language models, have emerged rapidly, showcasing extraordinary intelligence and driving a new wave of AI innovation [33]. In this era of large foundation models, we must consider two key questions:

Does the classic RL paradigm, as represented by visual RL, still retain research value and necessity? Additionally, what role does DA play in this evolving landscape?

Clarifying the roles of visual RL and DA in the Age of Foundation Models is important for two reasons. First, it ensures that this survey's content remains pertinent and valuable in the current AI landscape. Second, it offers meaningful guidance for the future development of visual RL and DA. This section will address these two questions systematically. Primarily, the classic RL paradigm, exemplified by visual RL, remains an indispensable component in achieving super-human decision-making intelligence. Currently, there are two primary approaches to leveraging foundation models in decision-making tasks.

1. The first approach leverages pre-trained multi-modal large models for high-level perception and planning, harnessing their comprehensive understanding and reasoning capabilities to enhance strategic decision-making in complex environments [81, 99, 124]. In this context, RL remains crucial for training the low-level control policy. This policy generates specific actions based on task-relevant features extracted by foundation models, enabling effective agent-environment interaction [114, 159]. Consequently, RL bridges the gap between the high-level perception and planning capabilities of foundation models and the concrete action execution required in real-world scenarios.
2. The second category of methods adopts the pre-training paradigm for data-driven offline policy training [95, 139, 160], drawing inspiration from the success of foundation models in other domains. In this paradigm, offline pre-training can be viewed as seeking an optimal initialization for online RL, while online fine-tuning is imperative for achieving high-level decision-making intelligence [79, 128]. This approach addresses two critical aspects. Firstly, due to the inherent distribution shift between offline data and real-world scenarios, online RL is necessary to correct these biases [201]. Secondly, only through exploration in online learning can the agent transcend the limitations of human-collected offline data and potentially surpass human-level intelligence [87, 175].

Furthermore, DA will undoubtedly continue to play a crucial role in visual RL and other RL tasks that utilize high-dimensional features as input. Primarily, DA remains a direct and effective method for expanding datasets and incorporating human prior knowledge [59, 216]. In decision-making tasks such as robotic control, the availability of training data is significantly more limited compared to language and vision tasks. Consequently, designing more powerful DA techniques is essential for training large decision models. Moreover, as discussed in Section 6.3.2, DA effectively mitigates plasticity loss during online RL training. Without DA, even when employing pre-trained visual encoders for feature extraction, visual RL algorithms would still suffer from catastrophic plasticity loss, impeding efficient training [112, 199]. As foundation models continue to evolve, RL is tasked with handling increasingly complex decision-making scenarios, and agents are required to possess continual fine-tuning capabilities. This evolution in the field underscores the critical need for agents to retain adaptive capabilities, emphasizing the continued relevance and importance of DA research and application [19].

6.6 Limitations of DA

While DA has demonstrated significant benefits in visual RL, as extensively discussed in the preceding sections, it is crucial to acknowledge its limitations and potential drawbacks to provide a comprehensive understanding of its applicability and effectiveness.

1. The applications of DA are highly task-specific and require extensive expert knowledge [141]. In practice, DA's effectiveness depends on prior knowledge of the variations between training and test environments, allowing for reliable specification of appropriate augmentation techniques [83]. For example, in DMControl-GB, only visual settings such as background colors are varied in the test environments, and specific DA techniques, such as random convolution, can effectively capture these prior variations [35, 55].
2. DA effectively mitigates plasticity loss in single-task visual RL, ensuring efficient training [112]. However, it proves insufficient for maintaining adequate plasticity in open-ended and continual RL scenarios, where significant plasticity degradation occurs over extended training periods despite its application [1, 2]. This limitation calls for more targeted interventions to maintain long-term plasticity in open-ended, multi-task RL, essential for developing adaptive foundational RL policies.
3. DA, which modifies observations after they are generated, provides versatile applicability without necessitating direct simulator manipulation [88]. This characteristic is particularly advantageous when the underlying simulation environment is inaccessible or unmodifiable. However, in scenarios where direct access to the simulator is available, domain randomization (DR) can generate more diverse and precise data [17, 118]. Consequently, in specific fields such as robot learning, DR may prove more effective than DA [66, 71].
4. Current DA techniques, primarily focused on image transformations of observations, are insufficient for generating truly diverse synthetic data. This limitation stems from their reliance on human-defined priors, which constrains the injection of novel, informative knowledge into the data [55]. However, the advent of increasingly powerful pre-trained generative models presents a promising avenue to overcome this bottleneck, potentially enabling DA to produce substantially richer training data [59].

7 Conclusion

In this paper, we present a comprehensive survey of DA in the paradigm of visual RL. We first propose the High-Dimensional Contextual Markov Decision Process

(HCMDP) as a general framework, elucidating the motivations for DA in improving sample efficiency and generalization. The main body of this survey meticulously examines existing related works, structured around two central themes: how to augment data and how to effectively utilize the augmented data. Subsequently, experimental results from widely used benchmarks demonstrate the efficacy of these techniques in visual RL. This survey also provides a list of current challenges and potential directions for future studies. In the following, we present a few suggestions and insights that are intended to benefit the relevant communities.

1. Compact and robust representation is vital for acquiring *sample-efficient* and *generalizable* visual RL agents; therefore, it is necessary to apply appropriate representation learning strategies to tackle the specific challenges of visual RL (Section 2.2). As a data-driven technique, DA is an essential component of representation learning and has great potential to be further explored (Section 2.3).
2. To fully harness the potential of DA, two complementary aspects must be addressed: how to augment data (Section 3) and how to effectively leverage augmented data (Section 4). The key to further advancing these aspects lies in two critical objectives: achieving semantic-level DA (Section 6.1), and attaining an effective balance between training stability and generalization ability (Section 6.2).
3. Beyond the common benefits such as regularization and feature manipulation that DA provides across all deep learning scenarios, there exists a unique mechanism behind DA's significant enhancement of visual RL training efficiency: its ability to effectively mitigate plasticity loss. Overcoming this RL-specific bottleneck should be a focal point for future research, emphasizing the development of more targeted augmentation strategies and other interventions (Section 6.3).
4. The difference between RL and SL should be given special attention when applying DA in visual RL, including fragile optimization process, the interactive data acquisition process and its absence of ground-truth labels (Section 6.4).
5. Visual RL, as a representative paradigm for learning control policies from high-dimensional features, remains a crucial component in the current era of large foundation models. DA, as a key element in achieving efficient and generalizable visual RL, warrants continued in-depth investigation (Section 6.5).

Overall, this survey strives to provide the first unified and principled framework for the large body of thriving research on DA in visual RL. We expect it to serve as a valuable guide for researchers and practitioners, and stimulate more inspiration in this fascinating field.

Acknowledgements This project is supported by the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001. Dr Tao's research is partially supported by NTU RSR and Start Up Grants.

References

1. Abbas Z, Zhao R, Modayil J, White A, Machado MC (2023) Loss of plasticity in continual deep reinforcement learning. In: Conference on Lifelong Learning Agents, PMLR, pp 620–636
2. Ahn H, Hyeon J, Oh Y, Hwang B, Moon T (2024) Reset & distill: A recipe for overcoming negative transfer in continual reinforcement learning. arXiv preprint arXiv:2403.05066
3. Almuzairee A, Hansen N, Christensen HI (2024) A recipe for unbounded data augmentation in visual reinforcement learning. arXiv preprint arXiv:2405.17416
4. Anand A, Racah E, Ozair S, Bengio Y, Côté MA, Hjelm RD (2019) Unsupervised state representation learning in atari. Advances in neural information processing systems
5. Antoniou A (2017) Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340
6. Ash, J., & Adams, R. P. (2020). On warm-starting neural network training. *Advances in neural information processing systems, 33*, 3884–3894.
7. Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research 3(Nov):397–422
8. Balestriero R, Bottou L, LeCun Y (2022a) The effects of regularization and data augmentation are class dependent. arXiv preprint arXiv:2204.03632
9. Balestriero R, Misra I, LeCun Y (2022b) A data-augmentation is worth a thousand samples: Exact quantification from analytical augmented sample moments. arXiv preprint arXiv:2202.08325
10. Beattie C, Leibo JZ, Teplyashin D, Ward T, Wainwright M, Küttler H, Lefrancq A, Green S, Valdés V, Sadik A, et al. (2016) Deepmind lab. arXiv preprint arXiv:1612.03801
11. Bellemare M, Dabney W, Dadashi R, Ali Taiga A, Castro PS, Le Roux N, Schuurmans D, Lattimore T, Lyle C (2019) A geometric perspective on optimal representations for reinforcement learning. Advances in neural information processing systems 32
12. Benton G, Finzi M, Izmailov P, Wilson AG (2020) Learning invariances in neural networks. Advances in Neural Information Processing Systems 2020
13. Bertoin D, Rachelson E (2022) Local feature swapping for generalization in reinforcement learning. arXiv preprint arXiv:2204.06355
14. Botev A, Bauer M, De S (2022) Regularising for invariance to data augmentation improves supervised learning. arXiv preprint arXiv:2203.03304
15. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.
16. Chen C, Li J, Han X, Liu X, Yu Y (2022) Compound domain generalization via meta-knowledge encoding. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7109–7119, https://doi.org/10.1109/CVPR52688.2022.00698
17. Chen X, Hu J, Jin C, Li L, Wang L (2021) Understanding domain randomization for sim-to-real transfer. arXiv preprint arXiv:2110.03239
18. Chen Z, Kiami S, Gupta A, Kumar V (2023) Genaug: Retargeting behaviors to unseen situations via generative augmentation. arXiv preprint arXiv:2302.06671

19. Chen Z, Mandi Z, Bharadhwaj H, Sharma M, Song S, Gupta A, Kumar V (2024) Semantically controllable augmentations for generalizable robot learning. arXiv preprint arXiv:2409.00951
20. Chepurko N, Marcus R, Zgraggen E, Fernandez RC, Kraska T, Karger D (2020) Arda: automatic relational data augmentation for machine learning. Proceedings of the VLDB Endowment
21. Cheung TH, Yeung DY (2020) Modals: Modality-agnostic automated data augmentation in the latent space. In: International Conference on Learning Representations
22. Choi J, Kim T, Kim C (2019) Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6830–6840
23. Cobbe K, Klimov O, Hesse C, Kim T, Schulman J (2019) Quantifying generalization in reinforcement learning. In: International Conference on Machine Learning, PMLR
24. Cobbe K, Hesse C, Hilton J, Schulman J (2020) Leveraging procedural generation to benchmark reinforcement learning. In: International conference on machine learning, PMLR
25. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV (2019) Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 113–123
26. Dablain DA, Chawla NV (2023) Towards understanding how data augmentation works with imbalanced data. arXiv preprint arXiv:2304.05895
27. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
28. DeVries T, Taylor GW (2017a) Dataset augmentation in feature space. arXiv preprint arXiv:1702.05538
29. DeVries T, Taylor GW (2017b) Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552
30. Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR), 27*(3), 326–327.
31. Doshi-Velez F, Konidaris G (2016) Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In: IJCAI: proceedings of the conference, NIH Public Access, vol 2016, p 1432
32. Du S, Krishnamurthy A, Jiang N, Agarwal A, Dudik M, Langford J (2019) Provably efficient rl with rich observations via latent state decoding. In: International Conference on Machine Learning, PMLR, pp 1665–1674
33. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. (2024) The llama 3 herd of models. arXiv preprint arXiv:2407.21783
34. Fan J, Li W (2022) Dribo: Robust deep reinforcement learning via multi-view information bottleneck. In: International Conference on Machine Learning, PMLR, pp 6074–6102
35. Fan L, Wang G, Huang DA, Yu Z, Fei-Fei L, Zhu Y, Anandkumar A (2021) Secant: Self-expert cloning for zero-shot generalization of visual policies. In: International Conference on Machine Learning
36. Farebrother J, Machado MC, Bowling M (2018) Generalization and regularization in dqn. arXiv preprint arXiv:1810.00123
37. Federici M, Dutta A, Forré P, Kushman N, Akata Z (2020) Learning robust representations via multi-view information bottleneck. arXiv preprint arXiv:2002.07017
38. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E (2021) A survey of data augmentation approaches for nlp. arXiv preprint arXiv:2105.03075
39. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning, PMLR, pp 1126–1135
40. François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J, et al. (2018) An introduction to deep reinforcement learning. Foundations and Trends® in Machine Learning 11(3-4):219–354
41. Ghosh D, Rahme J, Kumar A, Zhang A, Adams RP, Levine S (2021) Why generalization in rl is difficult: Epistemic pomdps and implicit partial observability. Advances in Neural Information Processing Systems 34
42. Gil, Y., Baek, J., Park, J., & Han, S. (2021). Automatic data augmentation by upper confidence bounds for deep reinforcement learning. *2021 21st International Conference on Control* (pp. 1199–1203). Automation and Systems (ICCAS): IEEE.
43. Golatkar AS, Achille A, Soatto S (2019) Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. Advances in Neural Information Processing Systems
44. Gong C, Wang D, Li M, Chandra V, Liu Q (2021) Keepaugment: A simple information-preserving data augmentation approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1055–1064
45. Greydanus S, Koul A, Dodge J, Fern A (2018) Visualizing and understanding atari agents. In: International conference on machine learning, PMLR, pp 1792–1801
46. Grigsby J, Qi Y (2020) Measuring visual generalization in continuous control from pixels. arXiv preprint arXiv:2010.06740
47. Guan L, Verma M, Guo S, Zhang R, Kambhampati S (2021) Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. Advances in Neural Information Processing Systems
48. Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International conference on machine learning, PMLR, pp 1861–1870
49. Hafner D, Lillicrap T, Ba J, Norouzi M (2019a) Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations
50. Hafner D, Lillicrap T, Fischer I, Villegas R, Ha D, Lee H, Davidson J (2019b) Learning latent dynamics for planning from pixels. In: International conference on machine learning, PMLR, pp 2555–2565
51. Hafner D, Lillicrap TP, Norouzi M, Ba J (2020) Mastering atari with discrete world models. In: International Conference on Learning Representations
52. Hallak A, Di Castro D, Mannor S (2015) Contextual markov decision processes. arXiv preprint arXiv:1502.02259
53. Han B, Zheng C, Chan H, Paster K, Zhang M, Ba J (2021) Learning domain invariant representations in goal-conditioned block mdps. Advances in Neural Information Processing Systems 34
54. Hansen N, Wang X (2021) Generalization in reinforcement learning by soft data augmentation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 13611–13617
55. Hansen N, Su H, Wang X (2021) Stabilizing deep q-learning with convnets and vision transformers under data augmentation. Advances in Neural Information Processing Systems 34
56. Hansen S, Dabney W, Barreto A, Van de Wiele T, Warde-Farley D, Mnih V (2019) Fast task inference with variational intrinsic successor features. arXiv preprint arXiv:1906.05030
57. Hausknecht M, Stone P (2015) Deep recurrent q-learning for partially observable mdps. In: 2015 aaai fall symposium series
58. Hausknecht M, Wagener N (2022) Consistent dropout for policy gradient reinforcement learning. arXiv preprint arXiv:2202.11818
59. He H, Bai C, Xu K, Yang Z, Zhang W, Wang D, Zhao B, Li X (2024) Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. Advances in neural information processing systems 36

60. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
61. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2021a) Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377
62. He T, Zhang Y, Ren K, Wang C, Zhang W, Li D, Yang Y (2021b) Aarl: Automated auxiliary loss for reinforcement learning. openreviewnet
63. He T, Zhang Y, Ren K, Liu M, Wang C, Zhang W, Yang Y, Li D (2022) Reinforcement learning with automated auxiliary loss search. arXiv preprint arXiv:2210.06041
64. Hendrycks D, Mu N, Cubuk ED, Zoph B, Gilmer J, Lakshminarayanan B (2019) Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781
65. Hernández-García A, König P (2018) Data augmentation instead of explicit regularization. arXiv preprint arXiv:1806.03852
66. Horváth, D., Erdős, G., Istenes, Z., Horváth, T., & Földi, S. (2022). Object detection using sim2real domain randomization for robotic applications. *IEEE Transactions on Robotics, 39*(2), 1225–1243.
67. Hu J, Jiang Y, Weng P (2024) Revisiting data augmentation in deep reinforcement learning. In: The Twelfth International Conference on Learning Representations,https://openreview.net/forum?id=EGQBpkIEuu
68. Huang T, Wang J, Chen X (2022) Accelerating representation learning with view-consistent dynamics in data-efficient reinforcement learning. arXiv preprint arXiv:2201.07016
69. Huang W, Yi M, Zhao X (2021) Towards the generalization of contrastive self-supervised learning. arXiv preprint arXiv:2111.00743
70. Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
71. Huber J, Hélénon F, Watrelot H, Amar FB, Doncieux S (2024) Domain randomization for sim2real transfer of automatically generated grasping datasets. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 4112–4118
72. Igl M, Ciosek K, Li Y, Tschiatschek S, Zhang C, Devlin S, Hofmann K (2019) Generalization in reinforcement learning with selective noise injection and information bottleneck. Advances in neural information processing systems 32
73. Immer A, van der Ouderaa TF, Fortuin V, Rätsch G, van der Wilk M (2022) Invariance learning in deep neural networks with differentiable laplace approximations. arXiv preprint arXiv:2202.10638
74. Imre B (2021) An investigation of generative replay in deep reinforcement learning. B.S. thesis, University of Twente
75. Kaiser L, Babaeizadeh M, Milos P, Osinski B, Campbell RH, Czechowski K, Erhan D, Finn C, Kozakowski P, Levine S, et al. (2019) Model-based reinforcement learning for atari. arXiv preprint arXiv:1903.00374
76. Kalashnikov D, Irpan A, Pastor P, Ibarz J, Herzog A, Jang E, Quillen D, Holly E, Kalakrishnan M, Vanhoucke V, et al. (2018) Scalable deep reinforcement learning for vision-based robotic manipulation. In: Conference on Robot Learning, PMLR
77. Kemertas, M., & Aumentado-Armstrong, T. (2021). Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems, 34*, 4764–4777.
78. Khalifa NE, Loey M, Mirjalili S (2021) A comprehensive survey of recent trends in deep learning for digital images augmentation. Artificial Intelligence Review pp 1–27
79. Kim J, Park S, Levine S (2024a) Unsupervised-to-online reinforcement learning. arXiv preprint arXiv:2408.14785
80. Kim, M., Rho, K., Kim, Yd., & Jung, K. (2022). Action-driven contrastive representation for reinforcement learning. *Plos one, 17*(3), Article e0265456.
81. Kim MJ, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, Rafailov R, Foster E, Lam G, Sanketi P, et al. (2024b) Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246
82. Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, Pérez P (2021) Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems
83. Kirk R, Zhang A, Grefenstette E, Rocktäschel T (2021) A survey of generalisation in deep reinforcement learning. arXiv preprint arXiv:2111.09794
84. Ko B, Ok J (2021) Time matters in using data augmentation for vision-based deep reinforcement learning. arXiv preprint arXiv:2102.08581
85. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*, 1097–1105.
86. Kumar S, Marklund H, Van Roy B (2023) Maintaining plasticity via regenerative regularization. arXiv preprint arXiv:2308.11958
87. Lambert N, Wulfmeier M, Whitney W, Byravan A, Bloesch M, Dasagi V, Hertweck T, Riedmiller M (2022) The challenges of exploration for offline reinforcement learning. arXiv preprint arXiv:2201.11861
88. Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., & Srinivas, A. (2020). Reinforcement learning with augmented data. *Advances in neural information processing systems, 33*, 19884–19895.
89. Laskin M, Srinivas A, Abbeel P (2020b) Curl: Contrastive unsupervised representations for reinforcement learning. In: International Conference on Machine Learning, PMLR, pp 5639–5650
90. Laskin M, Yarats D, Liu H, Lee K, Zhan A, Lu K, Cang C, Pinto L, Abbeel P (2021) Urlb: Unsupervised reinforcement learning benchmark. arXiv preprint arXiv:2110.15191
91. Laskin M, Liu H, Peng XB, Yarats D, Rajeswaran A, Abbeel P (2022) Cic: Contrastive intrinsic control for unsupervised skill discovery. arXiv preprint arXiv:2202.00161
92. Lee, A. X., Nagabandi, A., Abbeel, P., & Levine, S. (2020). Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems, 33*, 741–752.
93. Lee K, Lee K, Shin J, Lee H (2019) Network randomization: A simple technique for generalization in deep reinforcement learning. arXiv preprint arXiv:1910.05396
94. Lee, K. H., Fischer, I., Liu, A., Guo, Y., Lee, H., Canny, J., & Guadarrama, S. (2020). Predictive information accelerates learning in rl. *Advances in Neural Information Processing Systems, 33*, 11890–11901.
95. Lee S, Seo Y, Lee K, Abbeel P, Shin J (2022) Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In: Conference on Robot Learning, PMLR, pp 1702–1712
96. Li L, Lyu J, Ma G, Wang Z, Yang Z, Li X, Li Z (2024a) Normalization enhances generalization in visual reinforcement learning. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp 1137–1146
97. Li X, Shang J, Das S, Ryoo MS (2022) Does self-supervised learning really improve reinforcement learning from pixels? arXiv preprint arXiv:2206.05266
98. Li Y, Hu G, Wang Y, Hospedales T, Robertson NM, Yang Y (2020) Differentiable automatic data augmentation. In: European Conference on Computer Vision, Springer, pp 580–595

99. Li Z, Lu Y, Mu Y, Qiao H (2024b) Cog-ga: A large language models-based generative agent for vision-language navigation in continuous environments. arXiv preprint arXiv:2409.02522
100. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971
101. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
102. Lin, Y., Huang, J., Zimmer, M., Guan, Y., Rojas, J., & Weng, P. (2020). Invariant transform experience replay: Data augmentation for deep reinforcement learning. *IEEE Robotics and Automation Letters, 5*(4), 6615–6622.
103. Liu H, Abbeel P (2021a) Aps: Active pretraining with successor features. In: International Conference on Machine Learning, PMLR, pp 6736–6747
104. Liu H, Abbeel P (2021b) Behavior from the void: Unsupervised active pre-training. Advances in Neural Information Processing Systems 34
105. Liu M, Zhu Y, Chen Y, Zhao D (2024) Enhancing reinforcement learning via transformer-based state predictive representations. IEEE Transactions on Artificial Intelligence
106. Liu X, Zou Y, Kong L, Diao Z, Yan J, Wang J, Li S, Jia P, You J (2018) Data augmentation via latent space interpolation for image classification. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 728–733
107. Liu Z, Li X, Kang B, Darrell T (2020) Regularization matters in policy optimization-an empirical study on continuous control. In: International Conference on Learning Representations
108. Lu C, Huang B, Wang K, Hernández-Lobato JM, Zhang K, Schölkopf B (2020) Sample-efficient reinforcement learning via counterfactual-based data augmentation. arXiv preprint arXiv:2012.09092
109. Lu C, Ball P, Teh YW, Parker-Holder J (2024) Synthetic experience replay. Advances in Neural Information Processing Systems 36
110. Lyle C, Zheng Z, Nikishin E, Pires BA, Pascanu R, Dabney W (2023) Understanding plasticity in neural networks. In: International Conference on Machine Learning, PMLR, pp 23190–23211
111. Lyle C, Zheng Z, Khetarpal K, Martens J, van Hasselt H, Pascanu R, Dabney W (2024) Normalization and effective learning rates in reinforcement learning. arXiv preprint arXiv:2407.01800
112. Ma G, Li L, Zhang S, Liu Z, Wang Z, Chen Y, Shen L, Wang X, Tao D (2024a) Revisiting plasticity in visual reinforcement learning: Data, modules and training stages. In: The Twelfth International Conference on Learning Representations, https://openreview.net/forum?id=0aR1s9YxoL
113. Ma G, Zhang L, Wang H, Li L, Wang Z, Wang Z, Shen L, Wang X, Tao D (2024b) Learning better with less: Effective augmentation for sample-efficient visual reinforcement learning. Advances in Neural Information Processing Systems 36
114. Ma Y, Song Z, Zhuang Y, Hao J, King I (2024c) A survey on vision-language-action models for embodied ai. arXiv preprint arXiv:2405.14093
115. Mandi Z, Bharadhwaj H, Moens V, Song S, Rajeswaran A, Kumar V (2022) Cacti: A framework for scalable multi-task multi-scene visual imitation learning. arXiv preprint arXiv:2212.05711
116. Mansourifar H, Chen L, Shi W (2019) Virtual big data for gan based data augmentation. In: 2019 IEEE International Conference on Big Data (Big Data), IEEE, pp 1478–1487
117. Mazoure B, Tachet des Combes R, Doan TL, Bachman P, Hjelm RD (2020) Deep reinforcement and infomax learning. In: Advances in Neural Information Processing Systems
118. Mehta B, Diaz M, Golemo F, Pal CJ, Paull L (2020) Active domain randomization. In: Conference on Robot Learning, PMLR, pp 1162–1176
119. Mei S, Misiakiewicz T, Montanari A (2021) Learning with invariances in random features and kernel models. In: Conference on Learning Theory, PMLR, pp 3351–3418
120. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. (2015) Human-level control through deep reinforcement learning. nature 518(7540):529–533
121. Moradi, R., Berangi, R., & Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review, 53*(6), 3947–3986.
122. Moreno-Barea FJ, Strazzera F, Jerez JM, Urda D, Franco L (2018) Forward noise adjustment scheme for data augmentation. In: 2018 IEEE symposium series on computational intelligence (SSCI), IEEE, pp 728–734
123. Mott A, Zoran D, Chrzanowski M, Wierstra D, Jimenez Rezende D (2019) Towards interpretable reinforcement learning using attention augmented agents. Advances in Neural Information Processing Systems 32
124. Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, Wang B, Dai J, Qiao Y, Luo P (2023) Embodiedgpt: Vision-language pre-training via embodied chain of thought. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 36, pp 25081–25094, https://proceedings.neurips.cc/paper_files/paper/2023/file/4ec43957eda1126ad4887995d05fae3b-Paper-Conference.pdf
125. Muratore F, Gruner T, Wiese F, Belousov B, Gienger M, Peters J (2022) Neural posterior domain randomization. In: Conference on Robot Learning, PMLR, pp 1532–1542
126. Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical imaging, 65*, 96–99.
127. Nair AV, Pong V, Dalal M, Bahl S, Lin S, Levine S (2018) Visual reinforcement learning with imagined goals. Advances in neural information processing systems 31
128. Nakamoto M, Zhai S, Singh A, Sobol Mark M, Ma Y, Finn C, Kumar A, Levine S (2024) Cal-ql: Calibrated offline rl pretraining for efficient online fine-tuning. Advances in Neural Information Processing Systems 36
129. Nguyen T, Luu TM, Vu T, Yoo CD (2021) Sample-efficient reinforcement learning representation learning with curiosity contrastive forward dynamics model. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 3471–3477
130. Ni T, Eysenbach B, Seyedsalehi E, Ma M, Gehring C, Mahajan A, Bacon PL (2024) Bridging state and history representations: Understanding self-predictive rl. arXiv preprint arXiv:2401.08898
131. Nikishin E, Schwarzer M, D'Oro P, Bacon PL, Courville A (2022) The primacy bias in deep reinforcement learning. In: International conference on machine learning, PMLR, pp 16828–16847
132. Nikishin E, Oh J, Ostrovski G, Lyle C, Pascanu R, Dabney W, Barreto A (2024) Deep reinforcement learning with plasticity injection. Advances in Neural Information Processing Systems 36
133. Nozawa K, Sato I (2022) Empirical evaluation and theoretical analysis for representation learning: A survey. arXiv preprint arXiv:2204.08226
134. Ohno, H. (2020). Auto-encoder-based generative models for data augmentation on regression problems. *Soft Computing, 24*(11), 7999–8009.
135. Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

136. Pan F, Zhang T, Luo L, He J, Liu S (2022) Learn continuously, act discretely: Hybrid action-space reinforcement learning for optimal execution. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), pp 3912–3918
137. Parisi S, Rajeswaran A, Purushwalkam S, Gupta A (2022) The unsurprising effectiveness of pre-trained vision models for control. In: International Conference on Machine Learning, PMLR, pp 17359–17371
138. Peng XB, Andrychowicz M, Zaremba W, Abbeel P (2018) Sim-to-real transfer of robotic control with dynamics randomization. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 3803–3810
139. Prudencio RF, Maximo MR, Colombini EL (2023) A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks and Learning Systems
140. Rafiee B, Jin J, Luo J, White A (2022) What makes useful auxiliary tasks in reinforcement learning: investigating the effect of the target policy. arXiv preprint arXiv:2204.00565
141. Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., & Fergus, R. (2021). Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems, 34*, 5402–5415.
142. Rakelly, K., Gupta, A., Florensa, C., & Levine, S. (2021). Which mutual-information representation learning objectives are sufficient for control? *Advances in Neural Information Processing Systems, 34*, 26345–26357.
143. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347
144. Schwarzer M, Anand A, Goel R, Hjelm RD, Courville A, Bachman P (2020) Data-efficient reinforcement learning with self-predictive representations. In: International Conference on Learning Representations
145. Schwarzer M, Rajkumar N, Noukhovitch M, Anand A, Charlin L, Hjelm RD, Bachman P, Courville AC (2021) Pretraining representations for data-efficient reinforcement learning. Advances in Neural Information Processing Systems 34
146. Schwarzer M, Ceron JSO, Courville A, Bellemare MG, Agarwal R, Castro PS (2023) Bigger, better, faster: Human-level atari with human-level efficiency. In: International Conference on Machine Learning, PMLR, pp 30365–30380
147. Shah RM, Kumar V (2021) Rrl: Resnet as representation for reinforcement learning. In: International Conference on Machine Learning, PMLR, pp 9465–9476
148. Shen L, Yang L, Chen S, Yuan B, Wang X, Tao D, et al. (2022a) Penalized proximal policy optimization for safe reinforcement learning. arXiv preprint arXiv:2205.11814
149. Shen R, Bubeck S, Gunasekar S (2022b) Data augmentation as feature manipulation. In: International conference on machine learning, PMLR, pp 19773–19808
150. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data, 6*(1), 1–48.
151. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al. (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 362
152. Sodhani S, Meier F, Pineau J, Zhang A (2022) Block contextual mdps for continual learning. In: Learning for Dynamics and Control Conference, PMLR, pp 608–623
153. Sokar G, Agarwal R, Castro PS, Evci U (2023) The dormant neuron phenomenon in deep reinforcement learning. In: International Conference on Machine Learning, PMLR, pp 32145–32168
154. Song X, Jiang Y, Tu S, Du Y, Neyshabur B (2019) Observational overfitting in reinforcement learning. In: International Conference on Learning Representations
155. Stone A, Ramirez O, Konolige K, Jonschkowski R (2021) The distracting control suite–a challenging benchmark for reinforcement learning from pixels. arXiv preprint arXiv:2101.02722
156. Stooke A, Lee K, Abbeel P, Laskin M (2021) Decoupling representation learning from reinforcement learning. In: International Conference on Machine Learning, PMLR, pp 9870–9879
157. Sun C, Qian H, Miao C (2022) Cclf: A contrastive-curiosity-driven learning framework for sample-efficient reinforcement learning. arXiv preprint arXiv:2205.00943
158. Tan Z, Wang K, Wang X (2024) Implicit curriculum in procgen made explicit. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems, https://openreview.net/forum?id=nZB1FpXUU6
159. Tang G, Rajkumar S, Zhou Y, Walke HR, Levine S, Fang K (2024) Kalie: Fine-tuning vision-language models for open-world manipulation without robot data. arXiv preprint arXiv:2409.14066
160. Tarasov D, Kurenkov V, Nikulin A, Kolesnikov S (2024) Revisiting the minimalist approach to offline reinforcement learning. Advances in Neural Information Processing Systems 36
161. Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DdL, Budden D, Abdolmaleki A, Merel J, Lefrancq A, et al. (2018) Deepmind control suite. arXiv preprint arXiv:1801.00690
162. Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P (2017) Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE
163. Tomar M, Mishra UA, Zhang A, Taylor ME (2021) Learning representations for pixel-based control: What matters and why? arXiv preprint arXiv:2111.07775
164. Tschannen M, Djolonga J, Rubenstein PK, Gelly S, Lucic M (2019) On mutual information maximization for representation learning. arXiv preprint arXiv:1907.13625
165. Von Kügelgen J, Sharma Y, Gresele L, Brendel W, Schölkopf B, Besserve M, Locatello F (2021) Self-supervised learning with data augmentations provably isolates content from style. Advances in Neural Information Processing Systems 34
166. Wang C, Luo X, Ross K, Li D (2022a) Vrl3: A data-driven framework for visual deep reinforcement learning. arXiv preprint arXiv:2202.10324
167. Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvinick M (2016a) Learning to reinforcement learn. arXiv preprint arXiv:1611.05763
168. Wang K, Kang B, Shao J, Feng J (2020a) Improving generalization in reinforcement learning with mixture regularization. Advances in Neural Information Processing Systems
169. Wang Q, Yang J, Wang Y, Jin X, Zeng W, Yang X (2023) Making offline rl online: Collaborative world models for offline visual reinforcement learning. arXiv preprint arXiv:2305.15260
170. Wang X, Lian L, Yu SX (2021) Unsupervised visual attention and invariance for reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6677–6687
171. Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N (2016b) Dueling network architectures for deep reinforcement learning. In: International conference on machine learning, PMLR, pp 1995–2003
172. Wang, Z., Liu, L., & Tao, D. (2020). Deep streaming label learning. *International Conference on Machine Learning (ICML), 119*, 9963–9972.
173. Wang Z, Liu L, Duan Y, Kong Y, Tao D (2022b) Continual learning with lifelong vision transformer. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 171–181
174. Wang Z, Liu L, Duan Y, Tao D (2022c) Sin: Semantic inference network for few-shot streaming label learning. IEEE Transactions on Neural Networks and Learning Systems pp 1–14, https://doi.org/10.1109/TNNLS.2022.3162747
175. Wen X, Yu X, Yang R, Bai C, Wang Z (2023) Towards robust offline-to-online reinforcement learning via uncertainty and smoothness. arXiv preprint arXiv:2309.16973
176. Wen Z, Li Y (2021) Toward understanding the feature learning process of self-supervised contrastive learning. In: International Conference on Machine Learning, PMLR
177. Wong SC, Gatt A, Stamatescu V, McDonnell MD (2016) Understanding data augmentation for classification: when to warp? In: 2016 international conference on digital image computing: techniques and applications (DICTA), IEEE, pp 1–6
178. Wu K, Wu M, Chen Z, Xu Y, Li X (2022) Generalizing reinforcement learning through fusing self-supervised learning into intrinsic motivation. The 36th AAAI Conference on Artificial Intelligence (AAAI 2022)
179. Xiao T, Radosavovic I, Darrell T, Malik J (2022) Masked visual pre-training for motor control. arXiv preprint arXiv:2203.06173
180. Xingyi Yang iW (2024) Neural metamorphosis. ECCV
181. Xu G, Zheng R, Liang Y, Wang X, Yuan Z, Ji T, Luo Y, Liu X, Yuan J, Hua P, et al. (2023) Drm: Mastering visual reinforcement learning through dormant ratio minimization. arXiv preprint arXiv:2310.19668
182. Yan E, Huang Y (2021) Do cnns encode data augmentations? In: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8
183. Yang R, Wang J, Geng Z, Ye M, Ji S, Li B, Wu F (2022a) Learning task-relevant representations for generalization via characteristic functions of reward sequence distributions. arXiv preprint arXiv:2205.10218
184. Yang S, Dong Y, Ward R, Dhillon IS, Sanghavi S, Lei Q (2022b) Sample efficiency of data augmentation consistency regularization. arXiv preprint arXiv:2202.12230
185. Yang S, Xiao W, Zhang M, Guo S, Zhao J, Shen F (2022c) Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610
186. Yang X, Wang X (2023) Diffusion model as representation learner. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 18938–18949
187. Yang X, Ye J, Wang X (2022d) Factorizing knowledge in neural networks. In: European Conference on Computer Vision, Springer, pp 73–91
188. Yang, X., Zhou, D., Liu, S., Ye, J., & Wang, X. (2022). Deep model reassembly. *Advances in neural information processing systems, 35*, 25739–25753.
189. Yarats D, Kostrikov I, Fergus R (2020) Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In: International Conference on Learning Representations
190. Yarats D, Fergus R, Lazaric A, Pinto L (2021a) Mastering visual continuous control: Improved data-augmented reinforcement learning. In: International Conference on Learning Representations
191. Yarats D, Fergus R, Lazaric A, Pinto L (2021b) Reinforcement learning with prototypical representations. In: International Conference on Machine Learning, PMLR
192. Yarats D, Zhang A, Kostrikov I, Amos B, Pineau J, Fergus R (2021c) Improving sample efficiency in model-free reinforcement learning from images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35
193. You B, Arenz O, Chen Y, Peters J (2022) Integrating contrastive learning with dynamic models for reinforcement learning from images. Neurocomputing
194. Yu T, Lan C, Zeng W, Feng M, Zhang Z, Chen Z (2021) Playvirtual: Augmenting cycle-consistent virtual trajectories for reinforcement learning. Advances in Neural Information Processing Systems 34
195. Yu T, Zhang Z, Lan C, Chen Z, Lu Y (2022) Mask-based latent reconstruction for reinforcement learning. arXiv preprint arXiv:2201.12096
196. Yu T, Xiao T, Stone A, Tompson J, Brohan A, Wang S, Singh J, Tan C, Peralta J, Ichter B, et al. (2023) Scaling robot learning with semantically imagined experience. arXiv preprint arXiv:2302.11550
197. Yu Y (2018) Towards sample efficient reinforcement learning. In: IJCAI, pp 5739–5743
198. Yuan Z, Ma G, Mu Y, Xia B, Yuan B, Wang X, Luo P, Xu H (2022a) Don't touch what matters: Task-aware lipschitz data augmentationfor visual reinforcement learning. arXiv preprint arXiv:2202.09982
199. Yuan Z, Xue Z, Yuan B, Wang X, Wu Y, Gao Y, Xu H (2022b) Pre-trained image encoder for generalizable visual reinforcement learning. In: First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022
200. Yuan Z, Wei T, Cheng S, Zhang G, Chen Y, Xu H (2024) Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. arXiv preprint arXiv:2407.15815
201. Yue Y, Kang B, Ma X, Xu Z, Huang G, Yan S (2022) Boosting offline reinforcement learning via data rebalancing. arXiv preprint arXiv:2210.09241
202. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6023–6032
203. Zhang A, Wu Y, Pineau J (2018a) Natural environment benchmarks for reinforcement learning. arXiv preprint arXiv:1811.06032
204. Zhang A, Lyle C, Sodhani S, Filos A, Kwiatkowska M, Pineau J, Gal Y, Precup D (2020a) Invariant causal prediction for block mdps. In: International Conference on Machine Learning, PMLR, pp 11214–11224
205. Zhang A, McAllister R, Calandra R, Gal Y, Levine S (2020b) Learning invariant representations for reinforcement learning without reconstruction. arXiv preprint arXiv:2006.10742
206. Zhang C, Vinyals O, Munos R, Bengio S (2018b) A study on overfitting in deep reinforcement learning. arXiv preprint arXiv:1804.06893
207. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412
208. Zhang J, Ma K (2022) Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
209. Zhang L, Deng Z, Kawaguchi K, Ghorbani A, Zou J (2020c) How does mixup help with robustness and generalization? arXiv preprint arXiv:2010.04819
210. Zhang R, Torabi F, Guan L, Ballard DH, Stone P (2019) Leveraging human guidance for deep reinforcement learning tasks. arXiv preprint arXiv:1909.09906
211. Zheng C, Wu G, Li C (2024) Toward understanding generative data augmentation. Advances in Neural Information Processing Systems 36
212. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017) Random erasing data augmentation. arXiv preprint arXiv:1708.04896
213. Zhou K, Yang Y, Qiao Y, Xiang T (2020) Domain generalization with mixstyle. In: International Conference on Learning Representations

214. Zhu J, Xia Y, Wu L, Deng J, Zhou W, Qin T, Liu TY, Li H (2022) Masked contrastive representation learning for reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence
215. Zhu Y, Wong J, Mandlekar A, Martín-Martín R (2020) robosuite: A modular simulation framework and benchmark for robot learning. arXiv preprint arXiv:2009.12293
216. Zhu Z, Zhao H, He H, Zhong Y, Zhang S, Yu Y, Zhang W (2023) Diffusion models for reinforcement learning: A survey. arXiv preprint arXiv:2311.01223
217. Zou D, Cao Y, Li Y, Gu Q (2023) The benefits of mixup for feature learning. In: International Conference on Machine Learning, PMLR, pp 43423–43479