How to Learn Domain-Invariant Representations for Visual Reinforcement Learning: An Information-Theoretical Perspective

Shuo Wang[1,2], **Zhihao Wu**[1,2], **Jinwen Wang**[1,2], **Xiaobo Hu**[1,2], **Youfang Lin**[1,2] and **Kai Lv**[1,2*]

[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing, China
{shuo.wang, zhwu, 20291281, xiaobohu, yflin, lvkai}@bjtu.edu.cn

Abstract

Despite the impressive success in visual control challenges, Visual Reinforcement Learning (VRL) policies have struggled to generalize to other scenarios. Existing works attempt to empirically improve the generalization capability, lacking theoretical support. In this work, we explore how to learn domain-invariant representations for VRL from an information-theoretical perspective. Specifically, we identify three Mutual Information (MI) terms. These terms highlight that a robust representation should preserve domain invariant information (return and dynamic transition) under significant observation perturbation. Furthermore, we relax the MI terms to derive three components for implementing a practical Mutual Information-based Invariant Representation (MIIR) algorithm for VRL. Extensive experiments demonstrate that MIIR achieves state-of-the-art generalization performance and the best sample efficiency in the DeepMind Control suite, Robotic Manipulation, and Carla.

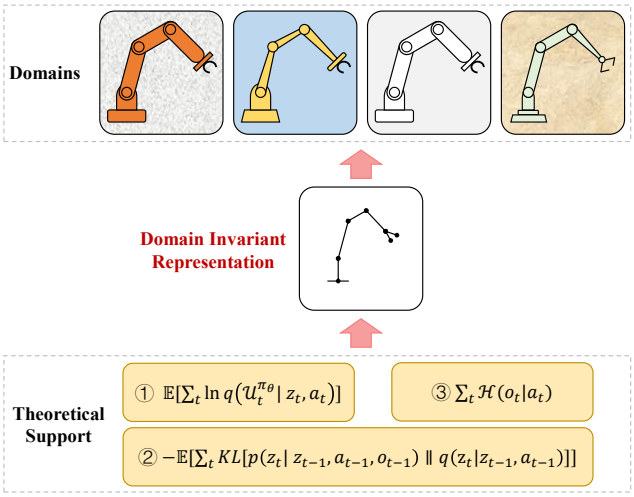

Figure 1: Illustration of our methods. We introduce information theory in visual reinforcement learning and identify three MI terms. Furthermore, we derive three theoretical supports for learning domain-invariant representations.

1 Introduction

Visual Reinforcement Learning (VRL) has achieved widespread success in various domains such as robotics control [Huang *et al.*, 2023], autonomous navigating [Wang *et al.*, 2023; Hu *et al.*, 2023], and video gaming [Zhou *et al.*, 2023]. However, learning a well-generalized policy across different domains in VRL remains challenging due to overfitting to irrelevant and redundant information.

Previous VRL methods can roughly be categorized into two classes: 1) introducing image processing techniques to capture the foreground or learn a consistent representation [Wang *et al.*, 2021; Hansen and Wang, 2021; Bertoin *et al.*, 2022]; and 2) exploiting the structural properties of RL to learn a robust policy, such as investigating data augmentation in TD update [Yarats *et al.*, 2021; Hansen *et al.*, 2021b]. Although prior studies have demonstrated empirical improvements, theoretical motivation and support remain lacking.

Different from previous works, we aim to introduce theoretical tools for VRL generalization. Information theory has garnered significant attention in deep learning, such as Information Bottleneck (IB) [Tishby and Zaslavsky, 2015]. IB believes that optimizing the Mutual Information (MI) between the representation and the input/output can lead to invariant representations with good generality. Then, [Xu and Raginsky, 2017] establishes a *theoretical link* between MI and the generalization capacities of deep learning. Nevertheless, utilizing information theory to explore generalization for VRL remains unexplored.

In this work, we introduce information theory to provide a novel perspective for VRL and identify three Mutual Information (MI) terms. Then, as depicted in Figure 1, we present three theoretical supports derived from the three MI terms for improving generalization. These theoretical supports are pivotal for acquiring domain-invariant representations, the sketch of the robotic arm, which is applicable across diverse domains. Specifically, these theoretical supports emphasize the necessity of preserving return and dynamic information under significant observation perturbation.

*Corresponding author: Kai Lv.

Based on the theoretical support, we propose a practical algorithm, Mutual Information-based Invariant Representation (MIIR). MIIR relaxes the MI terms and obtains three components: Q Predicting (QP), Dynamic Predicting (DP), and Deviation Perturbation Regularization (DPR). Specifically, QP maps various observations to the ground-truth Q values. DP is a dynamic transition model within the representation space independently of observations. DPR employs the pixel deviation within a batch to quantify the relevance of the action and perturbs the action-irrelevant pixels as much as possible.

We conduct extensive experiments across three benchmarks: DeepMind Control suite, Robotic Manipulation, and Carla (an autonomous driving environment). The agent is trained within a fixed background and tested in unseen environments with varying appearances. The results of numerous experiments demonstrate that MIIR has attained state-of-the-art generalization performance and sample efficiency.

Our primary contributions are summarized as follows:

- We introduce information theory into VRL to understand the generalization. Specifically, we identify three MI terms for learning domain-invariant representations.
- In practice, we relax the MI terms and present a Mutual Information-based Invariant Representation (MIIR) algorithm comprising Q Predicting, Dynamic Predicting, and Deviation Perturbation Regularization.
- We conduct experiments across three benchmarks, demonstrating that MIIR achieves state-of-the-art generalization performance and sample efficiency.

2 Related Work

2.1 Data Augmentation in Visual RL

Data augmentation is widely adopted in visual reinforcement learning to improve sample efficiency and generalization. RAD [Laskin *et al.*, 2020b] is the first study to investigate the influence of data augmentation in visual reinforcement learning. [Lee *et al.*, 2020] introduce Random Convolutions to alleviate overfitting to textures. SODA [Hansen and Wang, 2021] introduces a novel data augmentation, which utilizes additional images linearly added with observations. TLDA [Yuan *et al.*, 2022a] suggests not to augment the task-relevant pixels, which are measured by Lipschitz constants. SRM [Huang *et al.*, 2022] proposes Frequency-based data augmentation as an effective method to improve generalization. CG2A [Liu *et al.*, 2023] highlights the existence of gradient conflicts among various data augmentations and investigates how to integrate augmentation combinations better in visual RL. Unlike previous methods, we endeavor to perturb action-irrelevant pixels.

2.2 Generalization in Visual RL

VRL has become a research hot topic [Hafner *et al.*, 2020; Zhang *et al.*, 2021; Chebotar *et al.*, 2023; Han *et al.*, 2023] and numerous studies have investigated how to improve the generalization for VRL. Except for emphasizing data augmentation mentioned above, existing methods can be divided into two main paradigms: 1) methods based on visual processing and 2) methods focused on RL characteristics.

For visual processing, VAI [Wang *et al.*, 2021] focuses on extracting foreground through unsupervised keypoint detection and visual attention, which is computational. CURL [Laskin *et al.*, 2020a] introduces self-supervised learning to acquire consistent representations. SODA [Hansen and Wang, 2021] further incorporates a more advanced contrastive learning method, BYOL [Grill *et al.*, 2020]. PIE-G [Yuan *et al.*, 2022b] discuss that representations pre-trained on ImageNet can also contribute to the generalization. SGQN [Bertoin *et al.*, 2022] utilizes saliency maps to focus on critical information for decision-making.

For RL characteristics, DrQ [Yarats *et al.*, 2021] proposes augmenting both the target-Q network and Q network while optimizing TD loss. SECANT [Fan *et al.*, 2021] introduces a distillation method, separating the processes of representation and policy learning. PAD [Hansen *et al.*, 2021a] fine-tunes by leveraging the in-variance of inverse dynamics in the test environment. SVEA [Hansen *et al.*, 2021b] only augments the Q network for updating stably. TRP [Wang *et al.*, 2024] identifies policy divergence and Bellman error as critical factors impacting generalization. In this work, we focus on investigating VRL generalization from a theoretical perspective.

2.3 Information Theory in Generalization

Information theory is frequently employed to comprehend the representation process in deep learning [Tishby *et al.*, 2000]. Consequently, Information Bottleneck [Tishby and Zaslavsky, 2015] has received widespread attention. [Xu and Raginsky, 2017] derives an upper bound for the generalization error and the input-output mutual information. [Shwartz-Ziv and Tishby, 2017] visualizes the deep learning training process and demonstrates that IB has the ability to compress input for deep learning to obtain effective representation. This theoretical conclusion is widely utilized in image processing [Li *et al.*, 2022; Ahuja *et al.*, 2021], time series forecasting [Jia *et al.*, 2023], reinforcement learning [Eysenbach *et al.*, 2019], and other fields. In Reinforcement Learning, [Laskin *et al.*, 2022; Eysenbach *et al.*, 2019] utilize MI to quantify the difference between trajectories and acquire diverse skills. [Eysenbach *et al.*, 2021; Igl *et al.*, 2019] demonstrate that IB theory can assist agents in generalizing across different environments whose dynamic transitions differ from the training environment. Inspired by previous works, we introduce information theory in VRL for appearance generalization.

3 Preliminaries

3.1 Visual Reinforcement Learning

Reinforcement Learning aims to learn a decision-making policy by interacting with the environment [Sutton and Barto, 2018]. Visual reinforcement learning can be formulated as a Partially Observable Markov Decision Process (POMDP) $\mathcal{M} =< \mathcal{O}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, where $\mathcal{O}$ is the observation space, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ is the dynamic transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. The objective is to learn a policy π_θ that maximize the expected cumulative return $\mathcal{U}^{\pi_\theta} =$

$\mathbb{E}_{\boldsymbol{o}_t,\boldsymbol{a}_t\sim\tau}[\sum_{t=0}^{T-1}\gamma^t\mathcal{R}(\boldsymbol{o}_t,\boldsymbol{a}_t)]$, where τ represents the trajectories sampled by the policy π_θ.

In this work, we follow the previous works and utilize an off-policy paradigm that obtains optimal policy π^* by estimating state-action value function $Q_\phi(\boldsymbol{o}_t,\boldsymbol{a}_t): S\times A\mapsto\mathbb{R}$. Specifically, we minimize Temporal Difference (TD) error [Sutton, 1988] to update $Q_\phi(\boldsymbol{o}_t,\boldsymbol{a}_t)$, formulated as:

$$\mathcal{L}_{td}=\mathbb{E}_{\boldsymbol{o}_t,\boldsymbol{a}_t,\boldsymbol{o}_{t+1}\sim\mathcal{B}}[\frac{1}{2}\|Q_\phi(\boldsymbol{o}_t,\boldsymbol{a}_t)-y_t\|_2^2], \tag{1}$$

where $y_t=\mathcal{R}(\boldsymbol{o}_t,\boldsymbol{a}_t)+\gamma\mathbb{E}_{\pi_\theta}[Q^{\text{tgt}}(\boldsymbol{o}_{t+1},\boldsymbol{a}_{t+1})]$ and $\mathcal{B}$ is the replay buffer.

3.2 Generalization in Visual RL

In this work, we focus on generalization in Visual RL which is trained on a fixed environment and tested on unseen environments. Specifically, we consider a set of MDP $\mathbb{M}$. Each $\mathcal{M}_i\sim\mathbb{M}$ shares the same underlying structure except its unique observation space. Our purpose is to find a policy π that can maximize the expected cumulative return on $\mathcal{M}\sim\mathbb{M}$ in a zero-shot manner. Therefore, the objective of visual RL generalization can be formulated as:

$$\eta_{\mathcal{M}}(\pi)=\mathbb{E}_{\boldsymbol{o}_t,\boldsymbol{a}_t\sim(\pi,\mathcal{M})}[\sum_{t=0}^{T-1}\gamma^t\mathcal{R}(\boldsymbol{o}_t,\boldsymbol{a}_t)]. \tag{2}$$

For convenience, we denote the training environment as $\mathcal{M}_{\text{train}}$ and the test environments as $\mathcal{M}_{\text{test}}$. Since $\mathcal{M}_{\text{test}}$ can not obtained during training, we often utilize empirical risk at $\mathcal{M}_{\text{train}}$ to learn the policy. Therefore, the generalization error can be formulated as:

$$\mathcal{L}_{\text{gen}}=\eta_{\mathcal{M}_{\text{test}}}(\pi)-\eta_{\mathcal{M}_{\text{train}}}(\pi). \tag{3}$$

3.3 Information Bottleneck

Information Bottleneck (IB) [Tishby and Zaslavsky, 2015] has been widely adopted in deep supervised learning. The key opinion of IB is that an invariant representation Z needs to compress the input information X as much as possible while mapping the label Y, which can be formulated as:

$$\max_Z\mathcal{I}(Z;Y)-\lambda\mathcal{I}(Z;X). \tag{4}$$

After that, [Xu and Raginsky, 2017] theoretically establishes the relationship between generalization error and mutual information:

$$|\mathcal{L}_{\text{gen}}|\leq\sqrt{\frac{2\sigma^2}{n}\mathcal{I}(Z;X)}, \tag{5}$$

where σ is the output variance and n is the sample size. This upper bound suggests that controlling the mutual information between the input and output of an algorithm can control its generalization error.

4 Methodology

In this section, we first introduce information theory in visual RL. We identify three Mutual Information (MI) terms that contribute to the generalization. In practice, we relax the MI terms and propose a Mutual Information-based Invariant Representation (MIIR) with three components.

4.1 Information Theory in Visual RL

Equation (4) and Equation (5) show that well-generalized representations require controlling the input-output mutual information. Therefore, we can introduce the mutual information into the visual RL for learning domain-invariant representations with a certain policy π_θ:

$$\begin{aligned}&\max\mathcal{I}(\boldsymbol{z}_{1:T};\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{a}_{1:T})-\lambda\mathcal{I}(\boldsymbol{z}_{1:T};\boldsymbol{o}_{1:T})\\&=\max\underbrace{\mathcal{I}(\boldsymbol{z}_{1:T};\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{a}_{1:T})}_{\text{term1}}-\lambda\underbrace{\mathcal{I}(\boldsymbol{z}_{1:T};\boldsymbol{o}_{1:T}|\boldsymbol{a}_{1:T})}_{\text{term2}}\\&\quad-\lambda\underbrace{\mathcal{I}(\boldsymbol{o}_{1:T};\boldsymbol{a}_{1:T})}_{\text{term3}},\end{aligned} \tag{6}$$

where $\boldsymbol{z}$ is the representation.

The first term implies that the representation and cumulative return sequences should exhibit a robust correlation when given an action sequence. To derive this term, we utilize variational inference techniques [Alemi *et al.*, 2017]:

$$\begin{aligned}&\mathcal{I}(\boldsymbol{z}_{1:T};\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{a}_{1:T})\\&=\mathbb{E}[\ln p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{z}_{1:T},\boldsymbol{a}_{1:T})-\ln p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{a}_{1:T})]\\&\propto\mathbb{E}[\ln p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{z}_{1:T},\boldsymbol{a}_{1:T})]\\&\geq\mathbb{E}[\ln p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{z}_{1:T},\boldsymbol{a}_{1:T})]\\&\quad-KL[p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{z}_{1:T},\boldsymbol{a}_{1:T})||\prod_t q(\mathcal{U}^{\pi_\theta}_t|\boldsymbol{z}_t,\boldsymbol{a}_t)]\\&=\mathbb{E}[\sum_t\ln q(\mathcal{U}^{\pi_\theta}_t|\boldsymbol{z}_t,\boldsymbol{a}_t)].\end{aligned} \tag{7}$$

Therefore, we obtain a lower bound of term1 by adding a non-negatived KL divergence. Note that the marginal distribution $\ln p(\mathcal{U}^{\pi_\theta}_{1:T}|\boldsymbol{a}_{1:T})$ is constant since it is independent of representation.

The second term indicates that the representation sequence should exhibit maximal independence from the observation sequence when given the action sequence. We also utilize variational inference to derive an upper bound for term2:

$$\begin{aligned}&\mathcal{I}(\boldsymbol{z}_{1:T};\boldsymbol{o}_{1:T}|\boldsymbol{a}_{1:T})\\&=\mathbb{E}[\sum_t\ln p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1},\boldsymbol{o}_t)-\ln p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})]\\&=\mathbb{E}[\sum_t\ln p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1},\boldsymbol{o}_t)-\ln p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})\\&\quad+\ln q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})-\ln q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})]\\&=\mathbb{E}[\sum_t\ln p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1},\boldsymbol{o}_t)-\ln q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})]\\&\quad-KL[p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})||q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})]\\&\leq\mathbb{E}[\sum_t KL[p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1},\boldsymbol{o}_t)||q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})]].\end{aligned} \tag{8}$$

Note that Equation (8) introduces a variational approximation $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})$ to construct a solvable variational upper bound. This upper bound suggests that the representation should be able to characterize the dynamic transition function independently of observations.

The third term implies the necessity to decrease the dependence of actions on observations. We can express the term3

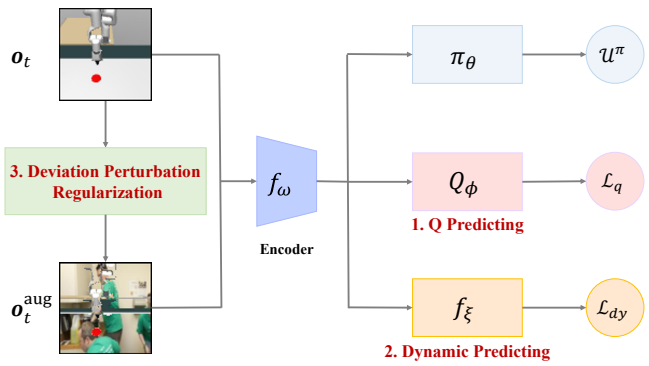

Figure 2: Overview. We first perturb action-irrelevant pixels through Deviation Perturbation Regularization. Then, observations $\boldsymbol{o}_t$ before and after augmentation are fed into multiple tasks: Q Predicting $\mathcal{L}_q$, Dynamic Predicting $\mathcal{L}_\xi$, and Policy Improvement $\mathcal{U}^{\pi_\theta}$. Only Policy Improvement utilizes original observations to stabilize policy updating. Notably, all tasks share the same encoder f_ω.

using the following formula:

$$\begin{aligned}&\mathcal{I}(\boldsymbol{o}_{1:T};\boldsymbol{a}_{1:T})\\&=\mathcal{H}(\boldsymbol{o}_{1:T})-\mathcal{H}(\boldsymbol{o}_{1:T}|\boldsymbol{a}_{1:T})\\&\leq-\mathcal{H}(\boldsymbol{o}_{1:T}|\boldsymbol{a}_{1:T})\\&=-\sum_t\mathcal{H}(\boldsymbol{o}_t|\boldsymbol{a}_t).\end{aligned}\tag{9}$$

$\mathcal{H}(\boldsymbol{o}_t|\boldsymbol{a}_t)$ reveals that learning domain-invariant representations requires increasing the entropy of observations without affecting the action.

In summary, we identify three theoretical supports from MI terms about how to learn domain-invariant representations. These objectives entail 1) maximizing the correlation between the representation and return, 2) preserving dynamic transition without depending on observations, and 3) increasing the entropy of the action-irrelevant pixels.

4.2 Mutual Information-based Invariant Representation

In this section, we propose Mutual Information-based Invariant Representation (MIIR) to learn domain-invariant representations. MIIR comprises three components: Q Predicting (QP), Dynamic Predicting (DP), and Deviation Perturbation Regularization (DPR). These components are derived by relaxing the above MI terms under certain mild assumptions. Figure 2 depicts the framework, and Algorithm 1 outlines the steps of our method.

Q Predicting Based on term1

Equation (7) shows that the representation $\boldsymbol{z}_t$ should retain sufficient information to predict the cumulative return $\mathcal{U}^{\pi_\theta}$ based on specific action sequences $\boldsymbol{a}_{1:T}$. Since the Q-function represents the expected cumulative return for taking a particular action in a given state, the Q-function and the expected cumulative return satisfy: $\mathcal{U}_t^{\pi_\theta}=\mathbb{E}_{\boldsymbol{a}_t\sim\pi}[Q_\phi^{\pi_\theta}(\boldsymbol{s}_t,\boldsymbol{a}_t)]$. Therefore, Equation (7) can be reformulated as:

$$\mathcal{I}(\boldsymbol{z}_{1:T};\mathcal{U}_{1:T}^{\pi_\theta}|\boldsymbol{a}_{1:T})\geq\mathbb{E}[\sum_t\ln q(Q_\phi^{\pi_\theta}(\boldsymbol{z}_t,\boldsymbol{a}_t))].\tag{10}$$

Furthermore, we assume that the output of $q(\cdot)$ follows a Gaussian distribution $\mathcal{N}(\mu_q,\sigma_q^2)$, where μ_q is the predicted mean of Q value and σ_q^2 is a pre-given variance. Therefore, the maximum likelihood objective in Equation (10) can be reformulated as a Mean Squared Error (MSE) loss, which is:

$$\mathcal{L}_q=\mathbb{E}[\|Q_\phi^{\pi_\theta}(\boldsymbol{z}_t,\boldsymbol{a}_t)-\sum_{h=0}^{H}\gamma^{t+h}R_{t+h}\|_2^2],\tag{11}$$

where representations $\boldsymbol{z}$ is embedded by a encoder $f_\omega(\cdot)$ from observations $\boldsymbol{o}$ or augmented observations $\mathrm{aug}(\boldsymbol{o})$.

Dynamic Predicting Based on term2

Equation (8) implies that a well-generalized representation can elucidate the dynamic transition function independently of observations. Specifically, predicting representation $\boldsymbol{z}_t$ only depended on last action $\boldsymbol{a}_{t-1}$ and representation $\boldsymbol{z}_{t-1}$. In practice, we employ a multivariate Gaussian distribution $\mathcal{N}(\mu_{\mathrm{dy}},\sigma_{\mathrm{dy}}^2)$ to characterize the output of the learned dynamic transition function. μ_{dy} represents the predicted mean of $\boldsymbol{z}_t$ and σ_{dy}^2 is a pre-given variance. Additionally, we assume that the dimensions of the multivariate Gaussian distribution are independent. Then, we can transform the KL divergence into MSE objective as:

$$\mathcal{L}_{\mathrm{dy}}=\mathbb{E}[\|f_\xi(\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})-f_\omega(\boldsymbol{o}_t)\|_2^2],\tag{12}$$

where $\boldsymbol{z}_t=f_\xi(\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})$ is the learned dynamic transition function. Note that $\boldsymbol{z}_{t-1}$ is embedded from observations $\boldsymbol{o}_{t-1}$ or augmented observations $\mathrm{aug}(\boldsymbol{o}_{t-1})$. We keep $\boldsymbol{o}_t$ unaugmented for stabilizing the learning process of $f_\xi(\boldsymbol{z}_{t-1},\boldsymbol{a}_{t-1})$.

Deviation Perturbation Regularization Based on term3

As shown in Equation (9), increasing the entropy of action-irrelevant pixels in observations is essential to acquire domain-invariant representations. We employ two mild assumptions [Misra *et al.*, 2024]: 1) action-relevant pixels exhibit a different frequency of change compared to others, and

Algorithm 1 Mutual Information-based Invariant Representation

ω,ϕ,ξ,θ: randomly initialized network parameters
α,β: learning rates
$\mathcal{B}$: empty replay buffer

for $t=1\cdots T$ **do**
  Sample action $\boldsymbol{a}_t\sim\pi_\theta(\boldsymbol{o}_t)$
  Interact with the environment $\boldsymbol{o}_{t+1}\sim\mathcal{P}(\boldsymbol{o}_t,\boldsymbol{a}_t)$
  Update transitions $\mathcal{B}\leftarrow\mathcal{B}\cup(\boldsymbol{o}_t,\boldsymbol{a}_t,R_t,\boldsymbol{o}_{t+1})$
  Sample a batch transitions $(\boldsymbol{o}_t,\boldsymbol{a}_t,R_t,\boldsymbol{o}_{t+1})$
  Perturb $\boldsymbol{o}_t$ as $\boldsymbol{o}_t^{\mathrm{aug}}$ with DPR
  if t mod update_frequency $=0$: **then**
    Update policy network $\theta\leftarrow\theta-\alpha\nabla_\theta\mathcal{U}^\pi(\theta)$
    Update Q Predicting $\phi\leftarrow\phi-\alpha\nabla_\phi\mathcal{L}_q(\phi)$
    Update Dynamic Predicting $\xi\leftarrow\xi-\beta\nabla_\xi\mathcal{L}_{dy}(\xi)$
  end if
end for

DMControl-GB (video-easy)	DrQ	SODA	TLDA	SVEA	PIE-G	SGQN	CG2A	MIIR
Walker, Walk	747 ± 21	768 ± 38	873 ± 34	819 ± 71	871 ± 22	910 ± 24	918 ± 20	$\mathbf{919 \pm 30}$
Walker, Stand	926 ± 30	955 ± 13	$\mathbf{973 \pm 6}$	961 ± 8	957 ± 12	955 ± 9	968 ± 6	971 ± 3
Ball_in_cup, Catch	380 ± 188	875 ± 56	887 ± 58	871 ± 106	922 ± 20	950 ± 24	963 ± 28	$\mathbf{973 \pm 2}$
Finger, Spin	599 ± 62	695 ± 97	744 ± 18	808 ± 33	837 ± 107	956 ± 26	912 ± 69	$\mathbf{978 \pm 9}$
Cartpole, Swingup	459 ± 81	758 ± 62	671 ± 57	782 ± 27	587 ± 61	761 ± 28	788 ± 24	$\mathbf{858 \pm 16}$
Cheetah, Run	102 ± 30	184 ± 64	356 ± 52	249 ± 20	287 ± 20	269 ± 33	314 ± 49	$\mathbf{393 \pm 57}$
DMControl-GB (video-hard)	DrQ	SODA	TLDA	SVEA	PIE-G	SGQN	CG2A	MIIR
Walker, Walk	121 ± 52	312 ± 32	271 ± 55	377 ± 93	600 ± 28	739 ± 21	687 ± 18	$\mathbf{821 \pm 58}$
Walker, Stand	252 ± 57	736 ± 132	602 ± 51	834 ± 46	852 ± 56	851 ± 24	895 ± 35	$\mathbf{965 \pm 2}$
Ball_in_cup, Catch	100 ± 40	381 ± 163	257 ± 57	403 ± 174	786 ± 47	782 ± 57	806 ± 44	$\mathbf{929 \pm 9}$
Finger, Spin	38 ± 13	309 ± 49	241 ± 29	335 ± 58	762 ± 59	822 ± 24	819 ± 38	$\mathbf{956 \pm 18}$
Cartpole, Swingup	136 ± 29	403 ± 17	286 ± 47	393 ± 45	401 ± 21	544 ± 43	472 ± 24	$\mathbf{765 \pm 22}$
Cheetah, Run	32 ± 13	94 ± 75	90 ± 27	105 ± 37	154 ± 17	144 ± 34	168 ± 16	$\mathbf{268 \pm 73}$

Table 1: Comparison with the SOTA on DMControl-GB. Our results reported include the average and standard deviation of return. Each result is calculated across 5 random seeds. We denote optimal performance in bold.

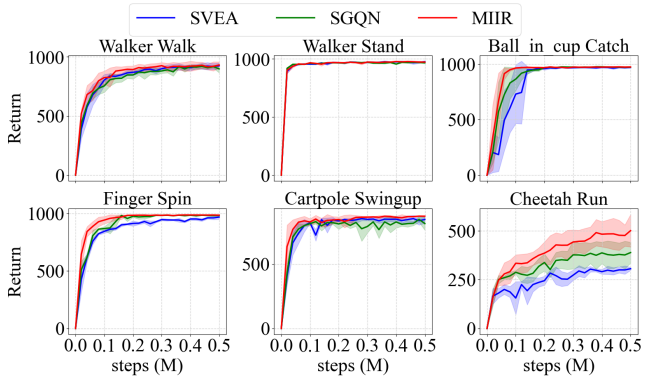

Figure 3: Sample efficiency in the training environment. We compare SVEA, SGQN, and MIIR across 6 tasks with 5 runs. MIIR (red) shows better asymptotic and sample efficiency during training.

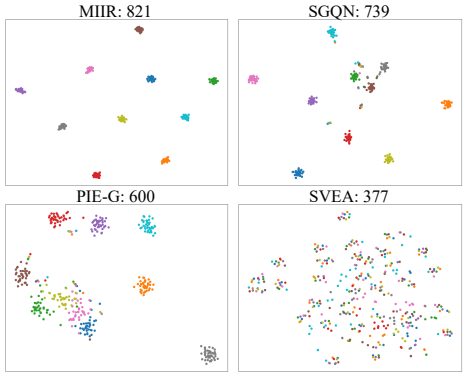

Figure 4: t-SNE of representation. We visualized the representation learned from MIIR, SGQN, PIE-G, and SEVA. We select 10 different states in **Walker walk** with 40 different backgrounds ("video_hard"). MIIR exhibits the best aggregation and separation.

2) the domain style remains *i.i.d.* throughout training. Therefore, pixels with high-frequency changing across transitions are salient and more likely to be action-relevant.

To measure the frequency of pixels, we introduce a batch kernel $\boldsymbol{k} \in \mathbb{R}^{N \times N \times B}$ inspired by the Mean Filter. N denotes the kernel size, and B represents the batch size of observations. First, we utilize the batch kernel to traverse all pixels and channels (with padding set to 1) to obtain a smooth feature map $\bar{\boldsymbol{o}} \in \mathbb{R}^{H \times W \times C}$. Notably, the smooth feature map has the same dimensions as the observation.

Then, we subtract the smooth feature map $\bar{\boldsymbol{o}}$ from the observations $\boldsymbol{o}$ to compute the deviation of each pixel. Specifically, a larger deviation indicates the pixel contains more high-frequency information which is more significant. To select more important pixels within each observation, we set a ρ threshold and obtain a binarized feature map M_ρ. $M_\rho(i,j) = 1$ if the pixel $\boldsymbol{o}(i,j)$ belongs to the ρ-quantile of highest deviation for $\boldsymbol{o}$, and 0 otherwise.

Finally, we perturb the pixels with lower deviation values by replacing observations to images $\widetilde{\boldsymbol{o}}$ sampled from Places [Zhou *et al.*, 2017]. Note that Places is a widely used dataset in Random Overlay (a data augmentation technique in visual RL). Therefore, DPR can be expressed as:

$$\mathrm{aug}(\boldsymbol{o}_t) = M_\rho \odot \boldsymbol{o}_t + (1 - M_\rho) \odot \widetilde{\boldsymbol{o}}_t, \tag{13}$$

where $\odot$ denotes the Hadamard product.

5 Experiments

In this section, we perform extensive experiments on DeepMind Control suite (DMControl) tasks [Tassa *et al.*, 2018], Robotic Manipulation tasks [Jangir *et al.*, 2022], and Carla [Dosovitskiy *et al.*, 2017] to evaluate generalization. We train the policy on a fixed environment and evaluate its generalization in environments with more complex backgrounds. The code has been released at: `https://github.com/Rebel-Uranus/MIIR`.

Task	Environment	SAC	SODA	SVEA	SGQN	CG2A	MIIR
Reach	Test1	-20.9 ± 16	-30.9 ± 43	-17.6 ± 10	14.4 ± 14	28.7 ± 1	$\mathbf{29.8 \pm 3}$
	Test2	-21.9 ± 14	-20.2 ± 29	-2.1 ± 39	31.0 ± 3	$\mathbf{36.7 \pm 4}$	31.9 ± 1
	Test3	-43.2 ± 6	-68.4 ± 30	1.4 ± 29	29.2 ± 7	$\mathbf{35.4 \pm 4}$	32.2 ± 2
	Test4	-36.7 ± 3	-37.9 ± 26	-22.2 ± 24	-5.6 ± 20	——	$\mathbf{29.0 \pm 2}$
	Test5	-49.9 ± 8	-52.3 ± 7	-33.4 ± 17	1.5 ± 15	——	$\mathbf{29.1 \pm 4}$
PegBox	Test1	-59.6 ± 26	16.9 ± 44	-21.3 ± 10	-72.1 ± 14	155.4 ± 17	$\mathbf{176.1 \pm 43}$
	Test2	-60.5 ± 12	0.7 ± 30	96.8 ± 41	110.7 ± 3	157.8 ± 22	$\mathbf{174.5 \pm 41}$
	Test3	-48.8 ± 17	73.6 ± 31	40.5 ± 28	154.6 ± 7	174.0 ± 21	$\mathbf{179.4 \pm 51}$
	Test4	-89.6 ± 31	-16 ± 83	33.2 ± 43	-70.9 ± 41	——	$\mathbf{163.8 \pm 46}$
	Test5	-75.7 ± 18	-100.9 ± 24	-98.7 ± 49	-87.4 ± 48	——	$\mathbf{182.2 \pm 42}$

Table 2: Comparison with the state-of-the-art methods on Robotic Manipulation tasks. Our results reported include the average and standard deviation of return. Each result is calculated across 5 random seeds. We denote the optimal result in bold. Since CG2A has not published its codes, we do not list the results on Test4 and Test5.

Baselines. We compare our method with current state-of-art approaches for generalization in visual reinforcement learning: 1) **DrQ** [Yarats *et al.*, 2021], a method adopts data augmentation for both Q-values and target Q-values when updating the TD loss; 2) **SODA** [Hansen and Wang, 2021], introducing BYOL [Grill *et al.*, 2020] and Random Overlay (a new data augmentation technique) to learn consistent representations; 3) **TLDA** [Yuan *et al.*, 2022a], an approach to avoid touching the action-relevant pixels when using data augmentation; 4) **SVEA** [Hansen *et al.*, 2021b], stabilizing Q learning by leaving target Q-values unaugmented during TD update; 5) **PIE-G** [Yuan *et al.*, 2022b], an algorithm that utilizes pre-trained ResNet [He *et al.*, 2016] representation into visual RL. 6) **SGQN** [Bertoin *et al.*, 2022], focusing on crucial pixels by introducing the saliency map; 7) **CG2A** [Liu *et al.*, 2023], alleviating gradient conflict when utilizes various data augmentations together.

5.1 DeepMind Control suite

We evaluate our algorithm on DMControl-GB [Hansen and Wang, 2021] across 6 tasks, *i.e.*, **Walker Walk, Walker Stand, Ball_in_cup Catch, Finger Spin, Cartpole Swingup**, and **Cheetah Run**. DMControl-GB provides *video-easy* and *video-hard* settings, where the backgrounds are replaced by various natural videos.

Generalization Performance. For each task, we run 5 seeds to compute the mean and standard deviation of return for evaluating generalization. As illustrated in Table 1, MIIR outperforms the baselines in 11 out of 12 cases regarding the return. Remarkably, our method attains the highest performance in the *video-hard* setting (the hardest test setting with complex video perturbation). In particular, MIIR obtains a 4% improvement compared with the second-best method in *video-easy* setting. Moreover, in *video-hard* setting, our algorithm improved the average return across all tasks by 22%. In conclusion, MIIR achieves remarkable generalization performance across all tasks in diverse test environments.

Sample Efficiency. To evaluate sample efficiency, we compare MIIR with SVEA and SGQN in the training domain across six tasks (each algorithm runs 5 seeds on each task). In Figure 3, the solid lines depict the average return, and the shaded regions represent one standard deviation. MIIR achieves the best sample efficiency and asymptotic performance across all tasks. Despite MIIR employing DPR to introduce noise into the observations, sample efficiency has not been adversely impacted. In particular, MIIR significantly improves in **Cheetah Run**.

Representation Visualization. To validate that MIIR can learn domain-invariant representations, we utilize t-SNE [Van der Maaten and Hinton, 2008] to visualize the learned features. First, we randomly select 10 different observations originating from different states and replace the background with 40 unseen images. Then, the 400 pre-processed images are embedded by the learned encoder of each algorithm. Observations from the same state are represented using the same color. Figure 4 demonstrates that algorithms with better invariant representations can achieve better generalization. Specifically, MIIR achieves the most effective aggregation for observations from the same state and optimal separation between different states.

5.2 Robotic Manipulation

To emphasize the generality of our method, we conduct experiments on Robotic Manipulation tasks introduced by [Jangir *et al.*, 2022]. The benchmark involves two goal-reaching tasks: 1) *Reach*, controlling a robotic arm to reach a red mask on the table; 2) *PegBox*, attempting to insert the peg into a box. Each task can be evaluated in 5 test environments. Only the texture or color of the background varies during testing. Since CG2A has not published the code, we utilize the results reported in their work. We compare our method with SAC, SODA, SVEA, SGQN, and CG2A.

As illustrated in Table 2, MIIR achieves the SOTA in 8 out of 10 test environments. In the *Reach* task, MIIR achieves the best performance in Test1, Test4, and Test5 environments. Specifically, MIIR demonstrates improvements of $6\times$ and $18\times$ in Test4 and Test5, respectively. In the *PegBox* task, MIIR outperforms other methods in all test environments. In particular, MIIR obtains $3.9\times$ and $3.4\times$ improvements in Test4 and Test5, respectively. Notably, our method achieves

DMControl-GB	QP	DP	DPR	Walker, Walk	Walker, Stand	Ball_in_cup, Catch	Finger, Spin	Cartpole, Swingup	Cheetah, Run
video_easy	✓		✓	892 ± 12	967 ± 2	971 ± 4	967 ± 27	842 ± 19	298 ± 30
		✓	✓	869 ± 43	969 ± 3	966 ± 3	970 ± 22	856 ± 5	388 ± 23
	✓	✓		889 ± 21	969 ± 3	916 ± 10	671 ± 49	844 ± 17	$\mathbf{405 \pm 26}$
	✓	✓	✓	$\mathbf{919 \pm 30}$	$\mathbf{971 \pm 3}$	$\mathbf{973 \pm 2}$	$\mathbf{978 \pm 9}$	$\mathbf{858 \pm 16}$	393 ± 57
video_hard	✓		✓	818 ± 42	944 ± 12	908 ± 24	917 ± 21	741 ± 39	179 ± 27
		✓	✓	795 ± 40	947 ± 12	904 ± 27	900 ± 50	635 ± 44	247 ± 37
	✓	✓		460 ± 67	846 ± 19	737 ± 30	281 ± 28	400 ± 27	103 ± 5
	✓	✓	✓	$\mathbf{821 \pm 58}$	$\mathbf{965 \pm 2}$	$\mathbf{929 \pm 9}$	$\mathbf{956 \pm 18}$	$\mathbf{765 \pm 22}$	$\mathbf{268 \pm 73}$

Table 3: Ablation study of different components. Results include the average and standard deviation of the return across 5 random seeds. We denote the optimal one in bold.

Weather	SVEA	SGQN	MIIR
HardRainNoon	199 ± 8	143 ± 30	$\mathbf{205 \pm 13}$
WetCloudySunset	187 ± 41	145 ± 32	$\mathbf{191 \pm 24}$
MidRainyNoon	162 ± 27	163 ± 8	$\mathbf{193 \pm 27}$
MidRainSunset	$\mathbf{212 \pm 15}$	156 ± 22	193 ± 24
WetNoon	77 ± 47	179 ± 13	$\mathbf{213 \pm 14}$
SoftRainNoon	192 ± 17	176 ± 13	$\mathbf{194 \pm 15}$
ClearNoon	197 ± 41	175 ± 10	$\mathbf{212 \pm 7}$
HardRainSunset	182 ± 14	163 ± 22	$\mathbf{217 \pm 31}$

Table 4: Comparison with SOTA on Carla. We demonstrate the average and standard deviation of return. Each result is calculated across 5 random seeds. We denote the optimal result in bold.

the highest return value with the smallest performance gap across various test environments on all tasks.

5.3 Autonomous Driving in Carla

We employ a more complex and realistic autonomous driving environment (Carla) to demonstrate the effectiveness of MIIR. The agent is trained on a "train" weather condition and tested in 8 unseen weather conditions. Each weather condition is characterized by a distinct combination of rain, clouds, wind, and sunshine. The goal is to steer the vehicle as far as possible within an episode. We run 5 seeds for each algorithm to calculate the average return and the standard deviation. Since the appearance change is mainly caused by different weathers, we adopt Random Convolution for SVEA and SGQN. In addition, the perturbation images $\widetilde{o}$ of MIIR are the observations o augmented by Random Convolution.

As shown in Table 4, MIIR achieves SOTA in 7 out of 8 test environments. Notably, MIIR exhibits stable performance across all test environments. On the contrary, SVEA and SGQN show more fluctuations in performance. SGQN exhibits weaker performance than SVEA and fails to maintain the advantages observed in the other benchmarks. The primary reason may be that SGQN needs to capture key pixels, which is challenging due to the complexity of Carla.

In conclusion, MIIR demonstrates the advancement across various tasks. Furthermore, MIIR exhibits high stability when generalized across different visual appearances.

5.4 Ablation Study

To illustrate the effectiveness of each component, we conduct experiments on DMC-GB, comparing different combinations of components. All combinations are trained for $500,000$ steps, evaluated across 6 tasks, and tested in "video_easy" and "video_hard" settings. Table 3 presents the results with one stand derivation.

We remove each component from MIIR to investigate its role, respectively. "QP & DPR" and "DP & DPR" is competitive with MIIR in both settings. However, "QP & DPR" and "DP & DPR" show obvious differences with MIIR on **Cheetah Run**. MIIR consistently outperforms, indicating that the return and dynamic transition information are both important. Notably, "QP & DPR" and "DP & DPR" demonstrate different advantages across different tasks.

Furthermore, we replace DPR with Random Overlay, denoted as "QP & DP", to investigate the importance of perturbing observations as much as possible. In "video_easy", "QP & DP" performs comparably with MIIR on most tasks. On the contrary, "QP & DP" drops dramatically in "video_hard". The disparity under different settings highlights the importance of increasing the entropy of observations. In other words, perturbing action-irrelevant pixels is crucial for generalizing to environments with significant interference.

Although the roles of components may overlap, experiments have shown that each component contributes differently. In brief, the three components work synergistically to promote the generalization of VRL.

6 Conclusion

We introduce information theory into Visual Reinforcement Learning (VRL) to explore how to learn domain-invariant representations. Specifically, we identify three MI terms and proposed a practical algorithm named Mutual Information-based Invariant Representation (MIIR). MIIR contains three components obtained by relaxing each corresponding MI term. We conduct extensive experiments to validate the generalization capabilities across the DeepMind Control suite, Robotic Manipulation, and autonomous driving Carla. Meanwhile, MIIR achieves the best sample efficiency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62206013).

References

[Ahuja *et al.*, 2021] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

[Alemi *et al.*, 2017] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *5th International Conference on Learning Representations, ICLR*, 2017.

[Bertoin *et al.*, 2022] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:30693–30706, 2022.

[Chebotar *et al.*, 2023] Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pages 3909–3928. PMLR, 2023.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[Eysenbach *et al.*, 2019] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations*, 2019.

[Eysenbach *et al.*, 2021] Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *International Conference on Learning Representations*, 2021.

[Fan *et al.*, 2021] Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Anima Anandkumar. Secant: Self-expert cloning for zero-shot generalization of visual policies. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Hafner *et al.*, 2020] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[Han *et al.*, 2023] Dong Han, Beni Mulyana, Vladimir Stankovic, and Samuel Cheng. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors*, 23(7):3762, 2023.

[Hansen and Wang, 2021] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *International Conference on Robotics and Automation*, 2021.

[Hansen *et al.*, 2021a] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Hansen *et al.*, 2021b] Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in Neural Information Processing Systems*, 34:3680–3693, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2023] Xiaobo Hu, Youfang Lin, HeHe Fan, Shuo Wang, Zhihao Wu, and Kai Lv. Building category graphs representation with spatial and temporal attention for visual navigation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[Huang *et al.*, 2022] Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. Spectrum random masking for generalization in image-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:20393–20406, 2022.

[Huang *et al.*, 2023] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[Igl *et al.*, 2019] Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

[Jangir *et al.*, 2022] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics Autom. Lett.*, 7(2):3046–3053, 2022.

[Jia *et al.*, 2023] Yuxin Jia, Youfang Lin, Xinyan Hao, Yan Lin, Shengnan Guo, and Huaiyu Wan. Witran: Water-wave information transmission and recurrent acceleration network for long-range time series forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Laskin *et al.*, 2020a] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

[Laskin *et al.*, 2020b] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[Laskin *et al.*, 2022] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, 35:34478–34491, 2022.

[Lee *et al.*, 2020] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. *8th International Conference on Learning Representations*, 2020.

[Li *et al.*, 2022] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7399–7407, 2022.

[Liu *et al.*, 2023] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23436–23446, 2023.

[Misra *et al.*, 2024] Dipendra Misra, Akanksha Saran, Tengyang Xie, Alex Lamb, and John Langford. Towards principled representation learning from videos for reinforcement learning. *12th International Conference on Learning Representations, ICLR*, 2024.

[Shwartz-Ziv and Tishby, 2017] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.

[Sutton, 1988] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

[Tassa *et al.*, 2018] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018.

[Tishby and Zaslavsky, 2015] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (itw)*, pages 1–5, 2015.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2021] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised visual attention and invariance for reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6677–6687, 2021.

[Wang *et al.*, 2023] Shuo Wang, Zhihao Wu, Xiaobo Hu, Youfang Lin, and Kai Lv. Skill-based hierarchical reinforcement learning for target visual navigation. *IEEE Transactions on Multimedia*, 2023.

[Wang *et al.*, 2024] Shuo Wang, Zhihao Wu, Xiaobo Hu, Jinwen Wang, Youfang Lin, and Kai Lv. What effects the generalization in visual reinforcement learning: policy consistency with truncated return prediction. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5590–5598, 2024.

[Xu and Raginsky, 2017] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

[Yarats *et al.*, 2021] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021.

[Yuan *et al.*, 2022a] Zhecheng Yuan, Guozheng Ma, Yao Mu, Bo Xia, Bo Yuan, Xueqian Wang, Ping Luo, and Huazhe Xu. Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 3702–3708. ijcai.org, 2022.

[Yuan *et al.*, 2022b] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.

[Zhang *et al.*, 2021] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[Zhou *et al.*, 2023] Bohan Zhou, Ke Li, Jiechuan Jiang, and Zongqing Lu. Learning from visual observation via offline pretrained state-to-go transformer. *arXiv preprint arXiv:2306.12860*, 2023.