Environment Agnostic Representation for Visual Reinforcement learning

Hyesong Choi[1], Hunsang Lee[2], Seongwon Jeong[1], Dongbo Min[1]†
[1]Ewha W. University [2]Hyundai Motor Company

Abstract

Generalization capability of vision-based deep reinforcement learning (RL) is indispensable to deal with dynamic environment changes that exist in visual observations. The high-dimensional space of the visual input, however, imposes challenges in adapting an agent to unseen environments. In this work, we propose Environment Agnostic Reinforcement learning (EAR), which is a compact framework for domain generalization of the visual deep RL. Environment-agnostic features (EAFs) are extracted by leveraging three novel objectives based on feature factorization, reconstruction, and episode-aware state shifting, so that policy learning is accomplished only with vital features. EAR is a simple single-stage method with a low model complexity and a fast inference time, ensuring a high reproducibility, while attaining state-of-the-art performance in the DeepMind Control Suite and DrawerWorld benchmarks. Code is available at: `https://github.com/doihye/EAR`.

1. Introduction

Deep reinforcement learning (RL) plays an important role in various fields such as robotic manipulation, video games, and autonomous navigation. Among them, RL approaches using visual observations [32, 26, 31, 7, 58, 1, 46] have achieved appreciable success in that dense and rich information can be obtained easily through the camera. Since real world changes dynamically at all times, the ability to generalize an agent against visual variations is indispensable in this field. However, the high-dimensional observation space of visual inputs [5, 41] imposes several challenges in adapting the agent to unseen environments [8, 4].

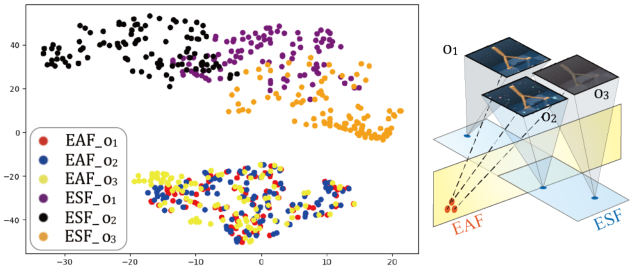

Figure 1. From visual observations with the same agent state under different environments (o_1: original environment, o_2: random box, o_3: color jitter), EAR separates environment-agnostic feature and environment-specific feature to learn vital representation used for RL policy learner.

To learn robust policies invariant to visual changes, domain randomization methods [44, 37, 36, 53, 38] have been proposed based on the surmise that applying miscellaneous augmentations during a training phase empowers an agent to cover any test environments. However, a considerable amount of effort is required to set the degree of randomization, and more importantly, it is practically impossible to fully consider all variations of test scenarios.

Instead of increasing the variability in training data, domain adaptation approaches have attempted to adapt the policy ingenuously to a test domain [39, 20, 19]. However, these methods often require additional fine-tuning during inference with a reward function that is newly created for each unseen environment. Alternatively, PAD [12] proposes to adapt an auxiliary self-supervised task (e.g. inverse dynamics prediction and rotation prediction) to obtain supervisory during test time without using reward. While this approach benefits greatly from an online learning, an appropriate auxiliary task should be selected depending on the specific RL task. More importantly, the additional training stage in the novel environment causes extra inference time, impairing an overall computational efficiency.

More recently, as an explicit approach based on domain generalization, VAI [49] proposes to learn visual foreground masks from augmented input images to feed only foreground-related information to a policy network. Though a high generalization performance was reported in VAI, several shortcomings still limit the use of this method. Since the policy training process requires two pre-training steps of auxiliary tasks (e.g. keypoint detection and visual attention prediction), the learning process is complex, requiring substantial computational costs compared to existing works. Moreover,

*This work was supported by the Mid-Career Researcher Program (2021R1A2C2011624) and the Basic Research Laboratory Program (2021R1A4A1032582) from NRF Korea, and the Artificial Intelligence Convergence Innovation Human Resources Development (RS-2022-00155966) from Institute of Information & communications Technology Planning & Evaluation (IITP). † Corresponding author: dbmin@ewha.ac.kr

the image that VAI finally uses for learning the policy is generated using the multiplication of the foreground mask and the augmented image, therefore remnants of the environmental changes (e.g. foreground changes) still remain, hampering the policy training.

In this paper, we propose a novel generalization method, termed **E**nvironment **A**gnostic **R**einforcement learning (**EAR**), which is a compact framework for domain generalization of the visual deep RL. To be more specific, our method attempts to extract an environment-agnostic feature (EAF) and use it to learn the RL policy network. By simply adopting novel objectives, our method can be trained in a single step, without the complicated procedure consisting of multiple individual training steps as in [49]. For extracting the EAFs from input images, we provide a feature factorization constraint to EAFs and remaining environment-specific features (ESFs) which are separated from the latents, and at the same time impose a reconstruction constraint computed by reversely combining EAFs and ESFs separated from two consecutive frames. Moreover, the proposed method introduces an episode-aware state shifting (ESS) constraint to EAFs in a self-supervised Siamese framework. By utilizing the proposed domain generalization task in an end-to-end manner with RL, EAR learns the process of extracting an EAF from an image, so that policy learning is accomplished only with the vital agent-related information, as depicted in Figure 1.

It is noteworthy that we aim to improve the generalization ability and at the same time maintain a lightweight encoder used by prior RL algorithms [11, 23] without any additional architectures [17, 51]. To achieve this, we assumed additive feature factorization and designed the novel objectives. Moreover, the ESS module imposes an important constraint that the separated EAFs, which we assume as a newly generated *state*, must be consistent in terms of the episode-aware state shifting. In Table 6, we show that the ESFs should be excluded when training the policy network. EAR is the first attempt to generate a novel agent-related states to make them episode-consistent, yielding a significant impact on the robustness of performance. This is easily confirmed by referring to the standard deviation of episode returns with and without the ESS module ($\mathcal{L}_{ss}$) in Table 5. The proposed ESS module works complementary to the feature factorization framework, since both constraints aim to encode agent-related representation.

In simulation, we present an extensive evaluation on the DeepMind (DM) Control Suite [43] including DM Control Generalization Benchmark [14] and Distracting Control Suite [42], which introduce a number of visual distractors to analyze whether the trained agent performs well in various environments. For evaluating the robustness of the trained agent on realistic textures, we also conduct experiments on the DrawerWorld [49] robotic manipulation tasks which add texture distortion and background distortion to MetaWorld [54] benchmark. Our method achieves a high generalization performance, outperforming the state-of-the-art methods in tasks of DM Control Suite [43] by up to 50.1% and DrawerWorld [49] by up to 97.7% with low model complexity and computational cost (Table 1). Notably, empirical evaluations show that while EAR does not require any extra costly adaptation during the test time [12] or using extra training stages [49], it achieves the state-of-the-art performance over existing methods.

To summarize, EAR has the following affirmative assets.
1) Novel domain generalization framework designed for visual RL. EAR proposes a novel framework tailored to RL setup, which segregates the intrinsic property needed for policy learning. Moreover, the generalization process of EAR requires no manual annotations or prior knowledge of environments. Accordingly, EAR can be readily adopted in any task that is subject to RL.
2) Simple single-stage training. The proposed method can be implemented by applying only novel objectives in a single training stage without several training steps.
3) Low complexity. The proposed architecture maintains a network size almost similar to that of the existing RL algorithms [11, 23], which do not consider the generalization ability of the agent. Also, the inference time is faster than other adaptation methods [12] or similar to other generalization methods since EAR does not require any adaptation during inference. See more details in Table 1.
4) Superior performance. Superior generalization performance of the proposed method was validated through extensive experiments on diverse test environments including DM Control Suite [43] and DrawerWorld [49].

2. Related Work

Self-supervised learning for RL. Self-supervised learning attempts to extract meaningful features from only unlabeled input images by defining a pretext task [45, 48, 56]. (*e.g.* rotation prediction [9], jig-saw puzzle solving [33], and context prediction [6]) or leveraging contrastive learning [2, 3, 10]. In recent years, the self-supervised learning has been actively leveraged in the field of RL. Inspired by standard self-supervised learning adopting auxiliary tasks without supervision, PAD [12] minimizes the RL and self-supervised objectives jointly to adapt a pretrained policy to an unseen environment with no reward. SPR [40] proposes self-predictive representation that predicts its own latent state representations. VAI [49] extracts visual foreground through unsupervised keypoint detection and visual attention to deliver an invariant visual feature to RL policy learner. Unlike leveraging self-supervised learning for adapting domain [12] or obtaining visual foreground [49], we explicitly learn environment-agnostic feature from an image with self-supervised feature factorization, so that policy learning is

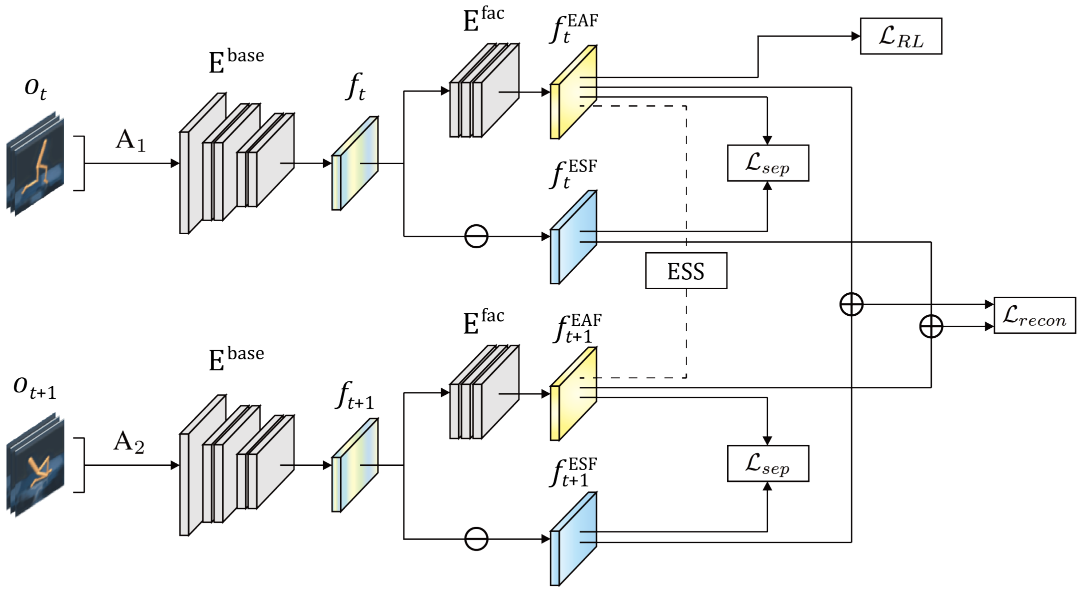

Figure 2. The overall framework of Environment Agnostic Reinforcement learning (EAR): EAR separates the environment-agnostic feature (EAF) from the latent representation and uses it to learn policies for RL. Through representation learning, we can guarantee that policy training proceeds only with the suitable feature from the observation of an unseen environment. A detailed description of the ESS module can be found in Figure 3.

accomplished only with the vital features.

Domain generalization for RL. Domain generalization for RL has received significant attention over the past few years. One approach is to enhance the robustness of policies against the visual changes. The use of randomly simulated RGB images was proposed in [44]. Similarly, Peng et al. [36] learn domain adaptive policies by randomizing the dynamics during a training phase. RARL [37] attempt to learn a robust policy against extra disturbances by modeling differences between training and test scenarios. Several works explore data augmentation techniques to improve the generalization capacity of policy [22, 38, 21]. For example, RAD [22] achieve a significant improvement via random translation and random amplitude scales, while DrAC [38] automatically find the most effective augmentation with regularization terms for the policy and value function. Rather than learning policies solely from augmented data, SODA [14] aims to decouple augmentation from policy learning by using non-augmented data for policy learning while using augmented data for auxiliary representation learning. More recently, SVEA [13] design a stabilized Q-value estimation framework for dealing with an instability issue under data augmentation in off-policy RL, while DBC [55] uses bisimulation metrics to learn a representation that disregards task-irrelevant information. Another promising approach is to adapt policy to a test domain. For example, Julian et al. [19] illustrate the effectiveness of fine-tuning in RL, while Rusu et al. [39] propose a progressive framework that accumulates prior knowledge while preventing catastrophic forgetting. Instead of directly adapting policy, PAD [12] proposed to adapt a self-supervised task to obtain free training signal during deployment. On the other hand, VAI [49] extracts a universal visual foreground mask to feed invariant observation to RL. For a similar purpose, but as a simpler and more effective way, our work focuses on generating an universal representation, which is invariant to distribution shifts, through novel objectives based on feature factorization, reconstruction, and episode-aware state-shifting.

Factorized representation. Feature factorizing, aiming to encode data into explanatory latent variables, has received considerable attention especially in image domain [24, 35, 27]. Recently, with the success of variants of variational autoencoders (VAE) [30, 29] and generative adversarial networks (GAN) [28, 34] even in the absence or lack of supervision, several attempts have been made to amalgamate factorized representation with RL. DARLA [17] and LUSR [51] exploit β-VAE [16] and cycle-consistent VAE [18] to encode an observation into environment's generative factors, respectively. However, the reconstruction loss of VAE-based approaches demands an extra decoder network and its implicit regularization still faces several difficulties in extracting environment-agnostic features from visual observation. On the other hand, the proposed feature factorizing method has the ability to factorize the visual observation into environment-agnostic and environment-specific features with the help of explicit objectives.

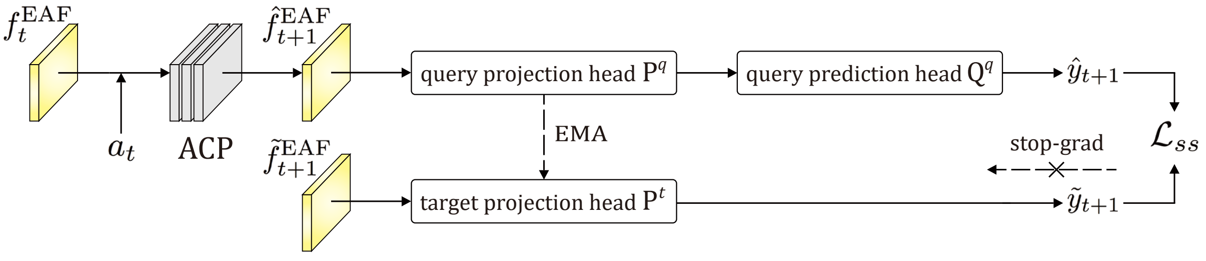
f_t^{EAF}
$\hat{f}_{t+1}^{\mathrm{EAF}}$
query projection head P^q
query prediction head Q^q
$\hat{y}_{t+1}$
a_t
ACP
EMA
$\mathcal{L}_{ss}$
$\tilde{f}_{t+1}^{\mathrm{EAF}}$
stop-grad
target projection head P^t
$\tilde{y}_{t+1}$

Figure 3. Episode-aware state shift (ESS) module: It consists of the action conditioned prediction (ACP), two projection heads, and one prediction head. The target projection head is updated via the exponential moving average (EMA) similar to self-supervised learning approaches.

3. Method

We propose Environment Agnostic Reinforcement learning (EAR), a compact and lightweight framework to generalize to an unseen environment in a completely unsupervised manner. EAR is an off-the-shelf method that can be readily adopted to any other RL algorithms. We cast the problem of generalization as an environment-agnostic feature extraction, which aims to separate the agent-related feature from input images, and use it to learn policies for RL.

3.1. Feature factorization of latent representation

As shown in Figure 2, EAR starts with two visual inputs with different kinds of augmentation A_1 and A_2 applied to suppose a new environment. The operation of EAR is regardless of the kind of augmentation as its ultimate purpose is to learn features in which environmental variations are removed. EAR learns feature factorization between the two groups with different augmentations in batch units, *i.e.*, the RL encoders for o_t and o_{t+1} are Siamese.

For encoders commonly used by off-the-shelf RL algorithms [23, 12, 22], we split the network into a base encoder $\mathrm{E}^{\mathrm{base}}$ and a factorization encoder $\mathrm{E}^{\mathrm{fac}}$. This allows the model complexity to remain unchanged by avoiding the use of additional layers for the factorization task. We found that such a compact architecture for the feature separation is very effective in terms of both accuracy and efficiency (see Table 1 and Table 2 and 3).

Given raw input observations o_t and o_{t+1} where t and $t+1$ represent a timestep, we first apply two different augmentations and obtain base features f_t and f_{t+1} which are the output of the base encoder $\mathrm{E}^{\mathrm{base}}$. Then, $\mathrm{E}^{\mathrm{fac}}$ separates EAF f_t^{EAF} and f_{t+1}^{EAF} from f_t and f_{t+1}, respectively. The difference between f_t and f_t^{EAF} is defined as the environment specific feature f_t^{ESF}.

$$\begin{aligned} f_t &= \mathrm{E}^{\mathrm{base}}(\mathrm{A}_1(o_t)), \\ f_t^{\mathrm{EAF}} &= \mathrm{E}^{\mathrm{fac}}(f_t), \\ f_t^{\mathrm{ESF}} &= f_t - f_t^{\mathrm{EAF}} \end{aligned} \tag{1}$$

This is similarly applied to f_{t+1}, where $f_{t+1} = \mathrm{E}_{base}(\mathrm{A}_2(o_{t+1}))$. Using the assumption that f_t^{EAF} and f_t^{ESF} must be independent and not related at all, we can impose a feature separation constraint which maximize the cosine distance between two features as follows:

$$\mathcal{L}_{sep} = \left| \left\| g(f_t^{\mathrm{EAF}}) \right\|_2^{\mathsf{T}} \left\| g(f_t^{\mathrm{ESF}}) \right\|_2 \right|, \tag{2}$$

where $g(\cdot)$ denotes a global average pooling operation to vectorize feature maps. Since subtraction between features is possible, conversely it is assumed that addition also holds true. In this regard, the addition of f_t^{EAF} and f_{t+1}^{ESF} should be formed similar to f'_t, which is the output feature map of E_{base} with the opposite augmentation applied, and it is likewise applied to f_{t+1}^{EAF} and f_t^{ESF} as well. We impose the mean square error (MSE) between features to give a reconstruction constraint $\mathcal{L}_{recon}$.

$$\begin{aligned} \mathcal{L}_{recon} = & \left\| (f_t^{\mathrm{EAF}} + f_{t+1}^{\mathrm{ESF}}) - f'_t \right\|_2^2 \\ & + \left\| (f_{t+1}^{\mathrm{EAF}} + f_t^{\mathrm{ESF}}) - f'_{t+1} \right\|_2^2, \end{aligned} \tag{3}$$

where $f'_t = \mathrm{E}_{base}(\mathrm{A}_2(o_t))$, and $f'_{t+1} = \mathrm{E}_{base}(\mathrm{A}_1(o_{t+1}))$. The environment-agnostic feature f_t^{EAF}, which is the output of $\mathrm{E}^{\mathrm{fac}}$, is finally used as an input for the RL policy network. In our work, the soft actor critic (SAC) [11] algorithm is used as a baseline.

3.2. Episode-aware state shift (ESS) module

The feature factorization and reconstruction constraints may be beneficial to separate input features, but an optimization using them alone may often lead to instability (refer to Table 5) due to the existence of multiple solutions satisfying these constraints. Besides, visual RL generally uses the image observation of the agent as an input, but disregards agent-related information.

Accordingly, we impose a novel complementary constraint that separated f^{EAF}, which we assume as a newly generated *state*, is consistent with respect to state shifting. We introduce the episode-aware state shift (ESS) module, which uses an action vector as a medium to constrain the factorized vital features of two consecutive frames. To be specific, the action conditioned prediction (ACP) first predicts EAF $\hat{f}^{\mathrm{EAF}}_{t+1}$ at the next frame from the current EAF f^{EAF}_{t} using the associated action a_t;

$$\hat{f}^{\mathrm{EAF}}_{t+1} = \mathrm{ACP}(f^{\mathrm{EAF}}_{t}, a_t). \tag{4}$$

Detailed structure of ACP is provided in the supplementary material. The state shift constraint is then imposed by applying two projection heads and one predictor as shown in Figure 3. We project the predicted representation $\hat{f}^{\mathrm{EAF}}_{t+1}$ and target representation $\tilde{f}^{\mathrm{EAF}}_{t+1}$ into a smaller latent space by feeding them into a query projection head P^q with parameters ξ^q and a target projection head P^t with parameters ξ^t. Note that, we follow the self-supervised literature [10] to compute the target feature $\tilde{f}^{\mathrm{EAF}}_{t+1}$ using the network updated via the exponential moving average (EMA). When using SAC [11] as the base algorithm, the feature of its critic target encoder can be reused as the target representation $\tilde{f}^{\mathrm{EAF}}_{t+1}$ for state shift constraint, instead of performing additional computation. Refer to the supplementary material for more implementation details about encoding target representation. After additionally applying the query prediction head Q^q to the query projection, the state shift objective $\mathcal{L}_{ss}$ is computed as follows:

$$\mathcal{L}_{ss}(\hat{y}_{t+1}, \tilde{y}_{t+1}) = -\frac{< \hat{y}_{t+1}, \tilde{y}_{t+1} >}{\|\hat{y}_{t+1}\|_2 \|\tilde{y}_{t+1}\|_2}, \tag{5}$$
$$\text{where} \quad \hat{y}_{t+1} = \mathrm{Q}^q(\mathrm{P}^q(\hat{f}^{\mathrm{EAF}}_{t+1})), \quad \tilde{y}_{t+1} = \mathrm{P}^t(\tilde{f}^{\mathrm{EAF}}_{t+1}).$$

The parameters of target projection head ξ_t are updated using the parameters of query projection head ξ_q via EMA, $\xi_t \leftarrow \tau\xi_t + (1-\tau)\xi_q$. Notably, the proposed state prediction works complementary to the feature factorization module, since it aims to learn state-related vital information. Namely, state representative feature contains vital information about the agent status, and it is essentially the same as EAFs, thus leading to mutual supplementation between two modules (feature factorization and ESS). Aggregated with the RL algorithm loss $\mathcal{L}_{RL}$, the final objective is as follows:

$$\mathcal{L}_{EAR} = \mathcal{L}_{RL} + \alpha\mathcal{L}_{sep} + \beta\mathcal{L}_{recon} + \gamma\mathcal{L}_{ss}. \tag{6}$$

4. Experiments

We intend to evaluate the generalization ability of the trained agent. Following [12, 14, 13, 49], the agent was first trained in the original environment and then evaluated under various environmental changes and visual distractors without any reward or prior knowledge about the test environment. For a comprehensive evaluation, we tested our method on both simulation and robotic manipulation. In the case of simulation, we presented an extensive evaluation on the challenging continuous control tasks of DM Control Suite [43] including DeepMind Control Generalization Benchmark [14] and Distracting Control Suite [42]. Beside simulations, we also conducted experiments on DrawerWorld [49] robotic manipulation tasks which is based on MetaWorld [54] benchmark with texture distortions and background distractions, in order to enable the agent to work in an environment close to real-world. The same RL algorithm (SAC [11]) was employed in our method and the existing methods used in the comparison. For reference, inspired by [15, 50, 52, 47], we employed an asymmetric augmentation intensity to the momentum encoder of SAC. Detailed experiment is provided in the supplementary material.

Table 1. Comparison on the number of model parameters and an average time per episode at evaluation time with VAI [49], SAC [11], PAD [12] and CURL [23]

Method	EAR	VAI	SAC	PAD	CURL
Model parameter (M)	2.55	4.88	2.54	2.54	2.54
Time per episode (s)	0.72	0.72	0.71	4.13	0.71

4.1. Model complexity and speed

Table 1 illustrates the comparison on the number of model parameters and an average speed per episode at evaluation time with VAI [49], SAC [11], PAD [12] and CURL [23]. EAR has the similar runtime and model complexity to the existing RL algorithms [11, 23] that do not consider the generalization capability. Even at training time, the increase in the number of parameters of the ESS module is very marginal at 0.5%, resulting in little changes to the overall model complexity. Compared to VAI [49] which requires twice as many parameters, and PAD [12] which takes more than four times for inference, EAR is much lighter and faster while achieving superior performance as reported in Section 4.2 and 4.3.

4.2. DeepMind Control Suite

Training. For tasks of DM Control Suite [43] 3D simulation benchmark, we trained agents on original environments without distractions. To ensure a fair comparison, we followed the environmental setting of current methods (PAD [12], SODA [14], SVEA [13], and VAI [49]). To guarantee reproducibility, we evaluated the model across 10 random seeds with 100 random environment initializations.

Test. Following prior works [12, 14, 13, 49], the generalization performance was measured on three different types of test environments: (i) randomized colors; (ii) video backgrounds; and (iii) camera poses as shown in Figure 4. For the test environments of randomized colors and video backgrounds, we used DM Control Generalization Benchmark [14]. We compared our method against the strong baselines including SAC [11] which is a base policy learning method, DR which is the SAC trained with simple domain randomization on a fixed set of 100 colors, PAD [12]

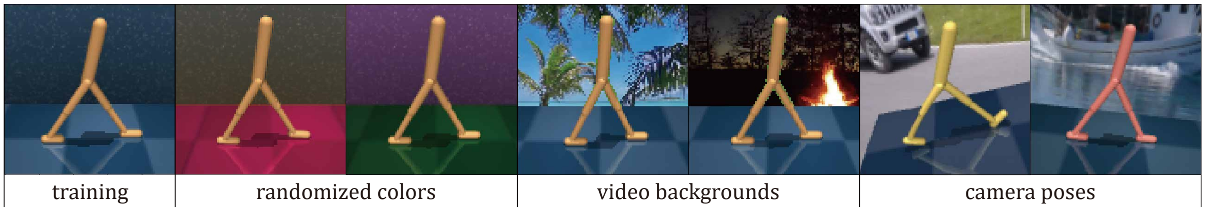

Figure 4. Training and test environments of DM Control Suite [43]. We train agents on original environment and evaluate them under various environmental changes and visual distractors without any reward or prior knowledge about the test environment in order to evaluate the generalization ability.

Table 2. Quantitative evaluation of episode return on DM Control [43] Generalization Benchmark [14] randomized color tests. EAR significantly outperforms existing state-of-the-art methods *without* any test time adaptation or using the additional training stages. We report our mean and standard deviation over 10 random seeds on 500K time steps. The results of other methods used for comparison are obtained from [12, 14, 13, 49]. The best result on each task is in bold.

Randomized colors	SAC	DR	PAD	SODA	SVEA	VAI	EAR	Δ
Walker, walk	414±74	594±104	468±47	692±68	760±145	819±11	**922±37**	+103
Walker, stand	719±74	715±96	797±46	893±12	942±26	964±2	**972±11**	+8
Cartpole, swingup	592±50	647±48	630±63	805±28	837±23	830±10	**884±39**	+47
Cartpole, balance	857±60	867±37	848±29	-	-	990±4	**997±2**	+7
Ball in cup, catch	411±183	470±252	563±50	949±19	961±7	886±33	**979±15**	+18
Finger, spin	626±163	465±314	803±72	793±128	**977±5**	932±3	946±18	-31
Finger, turn easy	270±43	167±26	304±46	-	-	445±36	**553±87**	+108
Cheetah, run	154±41	145±29	159±28	-	-	337±1	**368±56**	+31

Table 3. Quantitative evaluation of episode return on DM Control [43] Generalization Benchmark [14] video background tests. All experimental settings are the same as in Table 2.

Video backgrounds	SAC	DR	PAD	SODA	SVEA	VAI	EAR	Δ
Walker, walk	616±80	655±55	717±79	768±38	612±144	870±21	**913±38**	+43
Walker, stand	899±53	869±60	935±20	955±13	795±70	966±4	**970±23**	+4
Cartpole, swingup	375±90	485±67	521±76	758±62	606±85	624±146	**762±88**	+4
Cartpole, balance	693±109	766±92	687±58	-	-	869±189	**950±30**	+81
Ball in cup, catch	393±175	271±189	436±55	875±56	659±110	790±249	**911±40**	+36
Finger, spin	447±102	338±207	691±80	695±97	**764±86**	569±366	717±51	-47
Finger, turn easy	355±108	223±91	361±101	-	-	419±50	**629±39**	+210
Cheetah, run	194±30	150±34	206±34	-	-	322±35	**334±56**	+12

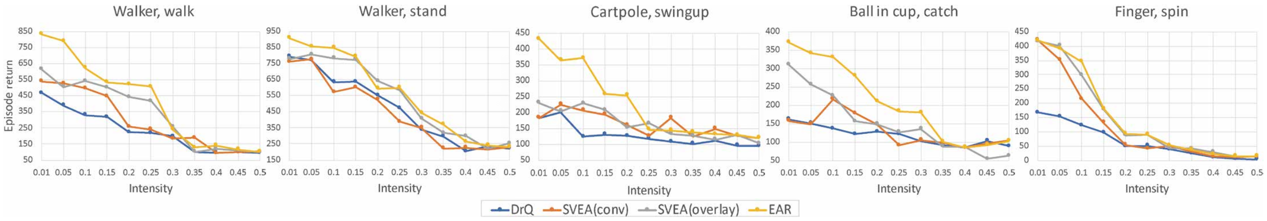

Figure 5. Quantitative evaluation of episode return on Distracting Control Suite [42] of DM Control Generalization Benchmark [14] camera pose tests.

which adapts to a test environment during deployment time, SODA [14] which applies soft data augmentation in an auxiliary learning, SVEA [13] which stabilizes Q-value estimation, and VAI [49] which extracts visual foreground mask. Note that, SODA [14] obtained the results using Places dataset [57] as a part of the training set, and SVEA [13] obtained the results with random convolution augmentation [25]. Additionally, we measured an episode return on Distracting Control Suite [42] of DM Control Generalization Benchmark [14], where camera pose, background, and colors continually change throughout an episode. Mean and standard deviation were measured over 10 random seeds on 500K time steps. We also provide evaluations on the sample efficiency in the supplementary material.

Randomized color test. Table 2 demonstrates that EAR outperforms the strong baselines on most of the tasks in terms of the episode reward return by up to 24.3%. When compared to PAD [12] tuning the encoder with test samples at test time, remarkable performance of EAR is possible even without an additional adaptation during deployment. This indicates the proposed generalization method is more efficient and effective than adaptation methods in terms of inference time

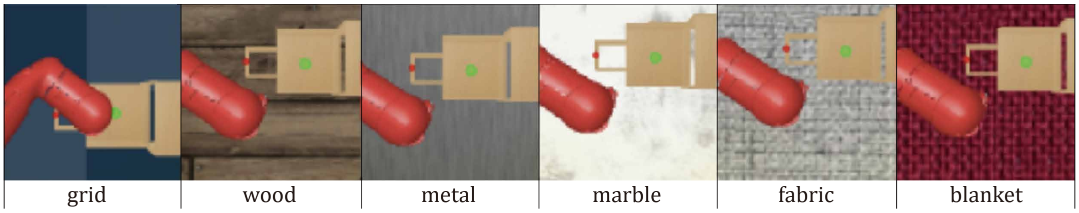

Figure 6. Training and test environments of DrawerWorld robotic manipulation benchmark [49] which is based on MetaWorld [54] benchmark with texture distortions and background distractions. We train agents on grid environment and evaluate them under texture distortions and background distractions without any reward or prior knowledge about the test environment in order to evaluate the generalization ability in terms of texture changes.

Table 4. Quantitative evaluation of episode return on DrawerWorld [49] robotic manipulation tasks. EAR surpasses strong baselines in most of the new texture environments of DrawerOpen and DrawerClose tasks. The mean and standard deviation were measured over 10 random seeds. The results of other methods used for comparison are obtained from [49]. The best result on each task is in bold.

success rate	DrawerOpen					DrawerClose				
	SAC	PAD	VAI	EAR	Δ	SAC	PAD	VAI	EAR	Δ
Grid	98±2	84±7	**100±0**	**100±0**	+0	**100±0**	95±3	99±1	**100±0**	+0
Black	95±2	95±3	100±1	**100±0**	+0	75±4	64±9	**100±0**	**100±1**	+0
Blanket	28±8	54±6	86±6	**92±4**	+6	0±0	0±0	85±8	**94±5**	+9
Fabric	2±1	20±6	**99±1**	93±6	-6	0±0	0±0	74±8	**90±8**	+16
Metal	35±7	81±3	98±2	**100±1**	+2	0±0	2±2	**98±3**	**98±1**	+0
Marble	3±1	3±1	43±7	**85±12**	+42	0±0	0±0	49±13	**77±16**	+28
Wood	18±5	39±9	94±4	**96±3**	+2	0±0	12±2	**70±6**	**70±8**	+0

(refer to Table 1) and accuracy. Meanwhile, it is noteworthy that even with a single training step, EAR derived greater performance than VAI [49] requiring complex plural training stages. These results indicate that using properly designed objectives to select vital environment-agnostic features is more crucial than adapting policies or deploying auxiliary networks with additional training stages.

Video background test. Table 3 shows that the proposed method greatly improves generalization ability over state-of-the-art baselines in 7 out of 8 tasks by up to 50.1%. Due to the nature of the proposed method of feeding environment-agnostic features to the RL algorithm, the trained agent come to be robust against a variety of challenging background distractions.

Camera pose test. Figure 5 represents that EAR competes favorably with the strong baselines on all the tasks of Distracting Control Suite [42] in terms of the randomization intensity. The intensity indicates the degree of variations, and some sample images with different intensities are provided in the supplementary material. This environment is much more challenging than other environments in that background, camera pose, and colors are constantly changing throughout a single episode. Nevertheless, since EAR utilizes features that eliminate environmental elements, it continuously performs well regardless of the environmental change. For low randomization intensity, EAR improved generalization by up to 87%, while the performance declined more slowly than those of the baselines as the intensity increased, ensuring high robustness.

4.3. DrawerWorld Robotics Manipulation

Training. For two tasks of DrawerWorld robotic manipulation benchmark [49], the agent was trained on the original environment without texture distortions and background distractions, which is referred to as 'Grid' in Table 4 and Figure 6. The DrawerOpen and DrawerClose tasks of DrawerWorld ask a Sawyer robot to open and close a drawer, and the reward function was set as a combination of a reaching reward and a push reward. To guarantee a fair comparison, we followed the environmental setting of [49].

Test. Following [49], the generalization performance was measured on six different types of new texture environments: Black, Blanket, Fabric, Metal, Marble and Wood including both color change and texture change of background, in order to enable the agent to work in a realistic environment. The performance was compared against SAC [11], PAD [12], and VAI [49]. We report mean and standard deviation of success rate which refers to the percentage of successful attempts out of 100 attempts over 10 random seeds.

Texture background test. Table 4 demonstrates that EAR remarkably outperforms the state-of-the-arts on the most tasks in terms of the success rate by up to 97.7%. The DrawerClose task, where the agent needs to infer a handle position precisely, is more challenging than the DrawerOpen task, and thus a success rate became 0 frequently in the existing approaches. However, EAR stably achieves a high success rate since the environment-agnostic features are fed to the

Table 5. Significance of three objectives on the DM Control [43] randomized color tests over 5 random seeds. The agent trained with a single objective produces an inferior performance to those of using two or more objectives, but it still produces comparable performance to state-of-the-arts [12, 14, 13, 49]. More details are found in the ablation study.

	Proposed objectives						
$\mathcal{L}_{sep}$	$\mathcal{L}_{recon}$	$\mathcal{L}_{ss}$	Walker, walk	Cartpole, swingup	Ball in cup, catch	Finger, spin	Cheetah, run
			412±68	528±63	545±47	757±49	149±54
✓			799±134	803±109	906±117	883±164	314±78
	✓		773±88	691±121	694±133	795±106	196±64
		✓	553±27	631±41	531±56	796±35	196±33
✓	✓		891±82	873±97	**979±104**	908±95	351±77
	✓	✓	873±49	839±34	952±29	927±38	288±21
✓		✓	879±50	856±28	955±67	912±44	352±42
✓	✓	✓	**922±37**	**884±39**	977±15	**946±18**	**368±56**

Table 6. Ablation study on where to apply the ESS constraint on the DM Control [43] randomized color tests. We showed that imposing the ESS module on the non-factorized original feature deteriorates the overall performance, due to the action-independence of environment-related information (ESF). This confirms that the ESS constraint should be imposed only on the environment-agnostic features (EAFs).

Application of state shift constraint	Walker, walk	Cartpole, swingup	Ball in cup, catch	Finger, spin	Cheetah, run
Applying non-factorized feature f_t to Eq. 5	919±43	805±24	968±12	935±32	**389±43**
Applying f_t^{EAF} to Eq. 5 (ours)	**922±37**	**884±39**	**977±15**	**946±18**	368±56

RL algorithm no matter what environmental change happens. Our results indicate that the visual features extracted by EAR is even robust to changes in background color and texture that are difficult to handle with CNNs.

4.4. Ablation Study

Significance of objectives. We conducted an intensive ablation study of the three objectives on the DM Control [43] randomized color tests over 5 random seeds in Table 5. The result without using all three objectives (first row) is the same as our baseline algorithm, SAC [11]. Our results can be summarized as follows. 1) The agent trained with a single objective produces an inferior performance to those of using two or more objectives, but it still produces comparable performance to state-of-the-arts [12, 14, 13, 49]. In particular, despite the relatively high standard deviation, the method with only $\mathcal{L}_{sep}$ applied produces much improved performance than the baseline, implying that environment-agnostic features can be learned to some extent with the simple feature separation constraint alone. 2) When $\mathcal{L}_{ss}$ was not used ($\mathcal{L}_{sep}$, $\mathcal{L}_{recon}$, $\mathcal{L}_{sep} + \mathcal{L}_{recon}$), the standard deviation was generally very high. This indicates that the episode-aware state shifting constraint enables stable generalization ability, since the objective imposes a constraint that the environment-agnostic feature should be consistent with respect to state shifting. 3) The performance gain by $\mathcal{L}_{sep}$ and $\mathcal{L}_{recon}$ was almost identical, in that both serve to separate the features. 4) Finally, the best performance was attained when the three objectives were used together. Refer to the supplementary material for more ablations.

Application of state shifting constraint. Table 6 provides the episode return depending on where the ESS constraint is applied on the DM Control [43] randomized color tests. Namely, we experimented the case of using the original non-factorized feature in Eq. 5 and compared it with the proposed method using EAFs. We showed that imposing the ESS module on the non-factorized feature deteriorates the overall performance by up to 9%. The performance degradation is attributed by the fact that environment-specific feature does not contain information related to the agent and therefore is not dependent on the action. Hence, it is ideal to apply the proposed ESS constraint to a representation that excludes the environment-specific information. In the proposed method, we supposed the separated EAFs as new state and imposed a state-shifting constraint as in Eq. 5.

5. Conclusion

We have presented a novel approach to address the essential generalization issue for visual RL. Our method is capable of consistently separating the environment-agnostic features from input visual observations with environment changes and feeding only them to the RL policy learner with the well-defined objectives considering the feature factorization, reconstruction, and episode-aware state shifting constraints. Empirical evaluations show that the proposed approach improves generalization over the state-of-the-arts on DM Control Suite and DrawerWorld benchmarks, while maintaining a low model complexity and a fast inference time.

Limitations and future work. While this work has showed superior performance on various benchmarks, similar to recent visual RL approaches, its applicability in real-world environments has not yet been fully validated.

References

[1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[4] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020.

[5] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pages 1282–1289. PMLR, 2019.

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[7] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[8] Shani Gamrian and Yoav Goldberg. Transfer learning for related reinforcement learning tasks via image-to-image translation. In *International Conference on Machine Learning*, pages 2063–2072. PMLR, 2019.

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[10] 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.

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[12] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *arXiv preprint arXiv:2007.04309*, 2020.

[13] 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, 2021.

[14] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13611–13617. IEEE, 2021.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[17] Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pages 1480–1490. PMLR, 2017.

[18] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018.

[19] Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning. *arXiv preprint arXiv:2004.10190*, 2020.

[20] D Kalashnikov, A Irpan, P Pastor, J Ibarz, A Herzog, E Jang, D Quillen, E Holly, M Kalakrishnan, V Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation (2018). *arXiv preprint arXiv:1806.10293*, 2018.

[21] Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

[22] 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.

[23] 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.

[24] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[25] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*, 2019.

[26] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[27] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *Advances in neural information processing systems*, 31, 2018.

[28] Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Detach and adapt: Learning cross-domain disentangled deep representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2018.

[29] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[30] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems*, 29, 2016.

[31] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[33] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[34] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[35] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pages 5102–5112. PMLR, 2019.

[36] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[37] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

[38] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[40] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

[41] Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning. *arXiv preprint arXiv:1912.02975*, 2019.

[42] Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite–a challenging benchmark for reinforcement learning from pixels. *arXiv preprint arXiv:2101.02722*, 2021.

[43] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[44] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 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)*, pages 23–30. IEEE, 2017.

[45] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[46] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[47] Xiao Wang, Haoqi Fan, Yuandong Tian, Daisuke Kihara, and Xinlei Chen. On the importance of asymmetry for siamese representation learning. *arXiv preprint arXiv:2204.00613*, 2022.

[48] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[49] 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.

[50] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *arXiv preprint arXiv:2104.07713*, 2021.

[51] Jinwei Xing, Takashi Nagata, Kexin Chen, Xinyun Zou, Emre Neftci, and Jeffrey L. Krichmar. Domain adaptation in reinforcement learning via latent unified state representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10452–10459, May 2021.

[52] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Hongkai Xiong, and Qi Tian. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *arXiv preprint arXiv:2012.02733*, 2020.

[53] Jiachen Yang, Brenden Petersen, Hongyuan Zha, and Daniel Faissol. Single episode policy transfer in reinforcement learning. *arXiv preprint arXiv:1910.07719*, 2019.

[54] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A

benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100. PMLR, 2020.

[55] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[57] 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.

[58] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.