ReCoRe: Regularized Contrastive Representation Learning of World Model

Rudra P.K. Poudel[1] Harit Pandya[1] Stephan Liwicki[1] Roberto Cipolla[1,2]
[1] Cambridge Research Laboratory
Toshiba Europe Ltd, UK
first-name.last-name@toshiba.eu
[2] Department of Engineering
University of Cambridge, UK
rc10001@cam.ac.uk

Abstract

While recent model-free Reinforcement Learning (RL) methods have demonstrated human-level effectiveness in gaming environments, their success in everyday tasks like visual navigation has been limited, particularly under significant appearance variations. This limitation arises from (i) poor sample efficiency and (ii) over-fitting to training scenarios. To address these challenges, we present a world model that learns invariant features using (i) contrastive unsupervised learning and (ii) an intervention-invariant regularizer. Learning an explicit representation of the world dynamics i.e. a world model, improves sample efficiency while contrastive learning implicitly enforces learning of invariant features, which improves generalization. However, the naïve integration of contrastive loss to world models is not good enough, as world-model-based RL methods independently optimize representation learning and agent policy. To overcome this issue, we propose an intervention-invariant regularizer in the form of an auxiliary task such as depth prediction, image denoising, image segmentation, etc., that explicitly enforces invariance to style interventions. Our method outperforms current state-of-the-art model-based and model-free RL methods and significantly improves on out-of-distribution point navigation tasks evaluated on the iGibson benchmark. With only visual observations, we further demonstrate that our approach outperforms recent language-guided foundation models for point navigation, which is essential for deployment on robots with limited computation capabilities. Finally, we demonstrate that our proposed model excels at the sim-to-real transfer of its perception module on the Gibson benchmark.

1. Introduction

In recent years, deep RL algorithms have been successfully employed for designing optimal strategies for games [25, 31] and shown a promise for controlling robots [21, 35]. Model-free RL approaches learn the policy along with the visual encoder in an end-to-end fashion from raw observations. As a result, they require a large number of training samples, which makes it difficult to deploy them on real robots where obtaining a large amount of training data is resource intensive, especially for safety-critical tasks such as autonomous navigation [9]. On the contrary, in model-based RL, an explicit predictive model of the world is learned called *world model*, enabling the agent to plan by thinking ahead [6, 10, 13, 31]. The world model is learned separately from the policy, therefore, the policy can use the world model as a surrogate for the real world. Consequently, model-based methods have higher sample efficiency [10, 13] making them more suitable in real environments since they can be trained with a small amount of data.

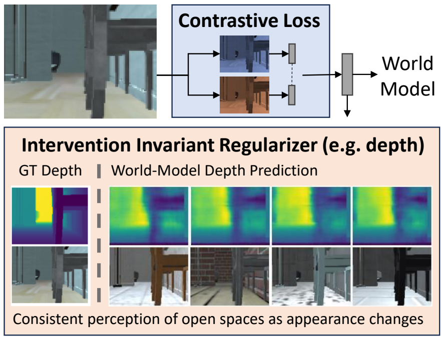
Contrastive Loss
World
Model
Intervention Invariant Regularizer (e.g. depth)
GT Depth
World-Model Depth Prediction
Consistent perception of open spaces as appearance changes

Figure 1. Intervention invariant regularizer is applied in addition to the contrastive loss in world-model-based RL. ReCoRe learns robust representations that are invariant to out-of-distribution appearance variations which help in the generalization of downstream tasks (such as navigation). Notice the consistent depth predictions despite texture variations of an iGibson evaluation scene.

Nevertheless, even current model-based RL struggles with generalization, as model-based approaches have to learn the world model purely from experience, which poses several challenges: The central issue is the training bias, which can be exploited by an agent, and leads to poor performance when deployed [10]. Another issue is that the latent representation is learned from a reconstruction loss, such as the state abstraction of variational autoencoders (VAE) [18], which is not sufficient to separate the task-relevant states from irrelevant ones. Hence, the RL policy may still overfit to environment-specific characteristics [40]. Thus, the aim of this paper is invariant feature abstraction, which is essential for learning a robust RL policy.

Specifically, we propose to use contrastive learning for invariant state abstraction since the contrastive learning objective implicitly ensures that the feature embeddings are invariant to the intervention, i.e. data augmentation. However, feature collapse is possible if a naïve implementation is used as can be seen in our evaluation (Table 1), where the naïve implementation of contrastive loss (ReCoRe-D) completely fails (<1% success rate) for model-based RL. Mitrovic *et al*. [24] proposed a regularizer based on KL-divergence that matches the distributions among the augmentations, which stabilizes contrastive learning under model-free RL settings and slightly improved the performance over CURL [20]. On the contrary, we propose a regularizer in the form of an auxiliary task to explicitly enforce the invariant feature learning. For example in the navigation task, we utilize depth predictions to extract the geometric features needed for navigation as they do not depend on textures as shown in Figure 1. We emphasize that depth is only required for training but not for deployment since it works as a regulariser. Furthermore, in cases where depth is not available other auxiliary tasks such as image denoising, segmentation or optical flow prediction can be utilized for regularization, enabling a wider applicability of the proposed model. Importantly, our setup allows us to employ contrastive learning in model-based RL settings, which improves the sample efficiency and helps with Out-of-Distribution (OoD) generalization.

In summary, we propose a *Regularized Contrastive Representation learning* (ReCoRe) approach to the world model. The proposed variant of the world model can extract and predict robust invariant features (Figure 2). ReCoRe is verified on the *point goal* navigation task from Gibson [37] and iGibson 1.0 [30] as well as on the DeepMind Control suite (DMControl) [33]. Thus, our main contributions are:

1. We show that contrastive unsupervised representation learning can significantly improve OoD generalization of world model based reinforcement learning (Table 1, ReCoRe vs. DreamerV2).
2. We propose an intervention-invariant regularizer that learns the invariant features (Section 3.1.1) and is shown to be crucial in preventing feature collapse of contrastive learning (Table 1, ReCoRe vs. ReCoRe-D).
3. Through extensive experiments, we showcase that our approach outperforms state-of-the-art RL models (including language guided foundation model Grounding DINO [23]) on out-of-distribution generalization (Table 1) and sim-to-real transfer of learned features (Table 2). We further show that even for in-distribution evaluation our approach outperforms model-free reinforcement learning approaches (Table 3) which is difficult for other model-based learning approaches.

2. Related Work

Unsupervised Representation Learning. Learning reusable feature representations from large unlabeled data is a fundamental challenge in machine learning. In the context of computer vision, one can leverage unlabeled images and videos to learn good intermediate representations, which can be useful for a wide variety of downstream tasks. Recently, VAE [18] has been a preferred approach for representation learning in model-based RL [10]. Since VAE does not make any additional consideration of downstream tasks, invariant representation learning with contrastive loss has shown more promising results [1, 20]. Self-supervised learning formulates representation learning as a supervised loss function between different transformations of data. In image-based learning self-supervision can be formulated using different image augmentations, for example, image distortion and rotation [3, 7]. We also use different data augmentation techniques to learn the invariant features using contrastive loss. Recently, transformer-based visual models such as DINO [2] have been shown to intrinsically capture robust object representations through self-supervision. Combining such models with language models pretrained on a large corpus of data [22, 23] have resulted in powerful representations that generalize well to diverse environments. However, inference on such large models is computationally expensive which makes it difficult to deploy on low-budget robots. We also compare our proposed approach to Grounding DINO [23] features, which we believe is the strongest baseline, and showcase superior results.

Contrastive Learning. Representation learning methods based on contrastive loss [4] have achieved state-of-the-art performance on face verification tasks. These methods use a contrastive loss to learn representations invariant to data augmentation [3, 16]. Given a list of input samples, contrastive loss forces samples from the same class to have similar embeddings and different ones for different classes. Since class labels are not available in the unsupervised setting, contrastive loss forces similar embedding for the augmented version of the same sample and different ones for different samples. There are several ways of formulating the contrastive loss such as Siamese [4], InfoNCE [34], and SimCLR [3]. In this work, we chose InfoNCE [34] for our contrastive loss

Learning Invariant Features. Learning structured representations that capture the underlying causal mechanisms generating the data is a central problem for robust machine learning systems [28]. However, recovering the underlying causal structure of the environment from observational

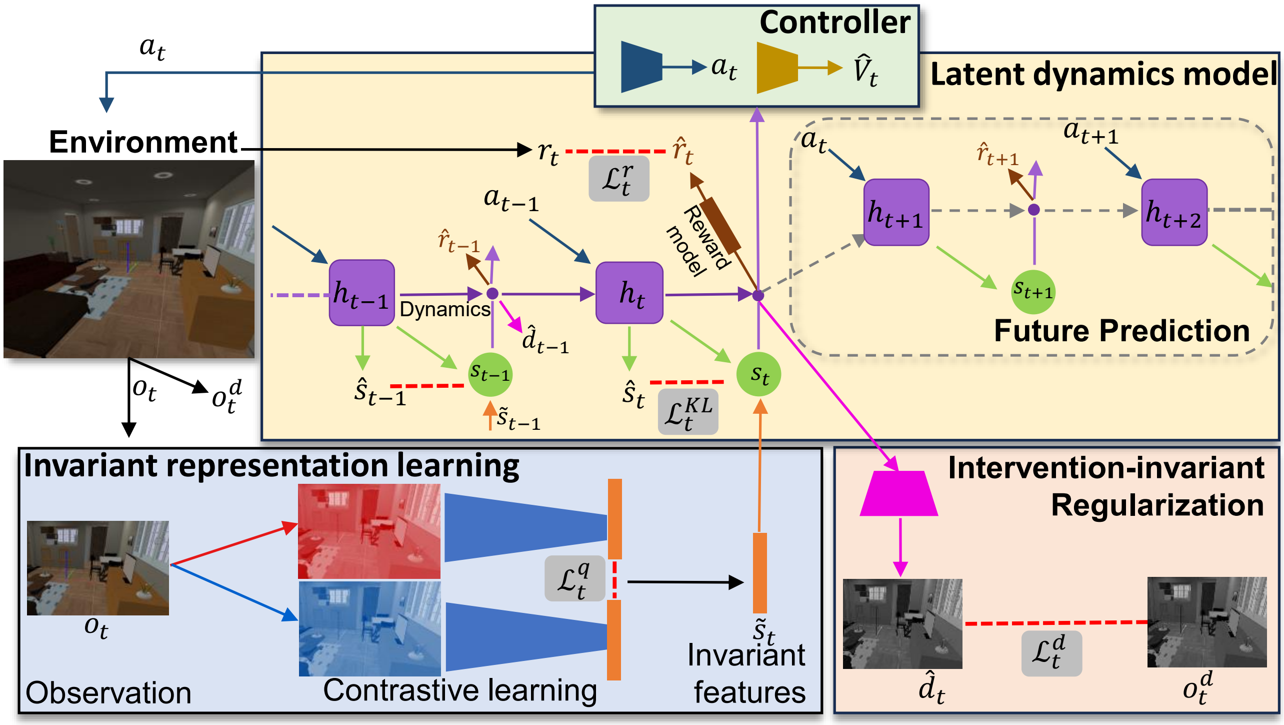

Figure 2. Flow diagram of proposed *Regularized Contrastive Representation learning* (ReCoRe) of World Model. It consists of four components: (i) invariant representation learning module, (ii) intervention-invariant regularizer, (iii) latent dynamics model, and (iv) actor-critic controller. The invariant representation learning module utilizes data augmentation and contrastive learning for invariant features abstraction ($\tilde{s}_t$) from image observations (o_t). The latent dynamics model employs a recurrent neural network with deterministic hidden states (h_t) to predict the stochastic latent prior states ($\widehat{s}_t$), and corresponding rewards ($\widehat{r}_t$) from the posterior (s_t). Intervention invariant regularizer considers an auxiliary task (here depth prediction i.e. $p_\theta(\widehat{d}_t|s_t, h_t)$) invariant to data augmentation (here texture variations) which prevents feature collapse in training the world-model with contrastive learning. The controller maximizes the expected rewards of the action using an actor critic approach. In addition to being sample efficient, the proposed approach is more robust to out-of-distribution and sim-to-real generalization, since the controller is learned separately using invariant states of the environment.

data without additional assumptions is a complex problem. A recent successful approach for causal discovery, in the context of unknown causal structure, is causal inference using invariant prediction [26]. Mitrovic *et al.* [24] recently formalized self-supervised representation learning using invariant causal mechanisms. Our proposed world model also exploits the invariance principle, which is formalized using contrastive loss to learn invariant features. In section 3, we explain how we utilized the data augmentation technique to learn the invariant state of the environment.

Model-based RL. The human brain discovers the underlying hidden causes of an observation. Those internal representations of the world influence how agents infer which actions will lead to a higher reward [15]. An early example of this idea was put forward by Sutton [32], where future hallucination samples rolled out from the learned world model are used in addition to the agent's interactions for sample efficient learning. Further, planning through the world model has been successfully demonstrated in the *world model* by Ha *et al.* [10] and DreamerV2 by Hafner *et al.* [13]. Recently, replacing state extraction [29] and dynamic prediction [27] using transformer architecture is a popular direction, which further improves the results [29]. Masked World Models (MWM) combines the transformer-based masked autoencoder with DreamerV2. Other Dreamer variants DayDreamer [36] and DreamerV3 [14] aim to scale up DreamerV2 architecture for physical robots and Atari games. Specifically, DayDreamer [36] trains DreamerV2 on physical robots, while DreamerV3 [14] proposes design choices (e.g. symlog scaling and EMA regularization) that help in scaling up DreamerV2 to several domains. In our work, we propose to learn invariant features to improve OoD generalization and sample efficiency further. While we utilize a similar architecture as DreamerV2 for policy and world model, our method is complementary to DreamerV3 and can take advantage of their design choices.

Sample Efficiency. Joint learning of auxiliary tasks with model-free RL makes them competitive with model-based

RL in terms of sample efficiency. For example, the recently proposed model-free RL method called CURL [20] added contrastive loss as an auxiliary task and outperformed the state-of-the-art model-based RL method called Dreamer [12]. Also, two recent works using data augmentation for RL called RAD [19] and DrQ [38] outperform CURL without using an auxiliary contrastive loss. These results warrant that if an agent has access to a rich stream of data from the environments, an additional regularizer is unnecessary since directly optimizing the policy objective is better than optimizing multiple objectives. However, we do not have access to a rich stream of data for many complex problems, hence sample efficiency still matters. Further, these papers do not consider the effect of regularizers in the form of auxiliary tasks and unsupervised representation learning for model-based RL, which is the main focus of our work.

3. ReCoRe-based World Model

We consider the visual control task as a finite-horizon partially observable Markov decision process (POMDP). We denote observation space, action space and time horizon as $\mathcal{O}$, $\mathcal{A}$ and $\mathcal{T}$ respectively. An agent performs continuous actions $a_t \sim p(a_t|o_{\leq t}, a_{<t})$, and receives observations and scalar rewards $o_t, r_t \sim p(o_t, r_t|o_{<t}, a_{<t})$ from the unknown environment. The goal of an agent is to maximize the expected total rewards $E_p(\sum_{t=1}^{T} r_t)$. In the following sections, we detail our proposed model.

3.1. World Model Design

We propose our *Regularized Contrastive Representation learning* (ReCoRe) technique to learn the world model. The data flow diagram of our proposed world model with explicitly regularized invariant feature learning is shown in Figure 2. Our method consists of four main components: (i) invariant representation learning module, (ii) intervention-invariant regularizer, (iii) latent dynamics model, and (iv) the controller. Next, we describe the components in detail.

3.1.1 Invariant Representations Learning Module

Extracting representations that are invariant to appearance variations from image observations is a key component of our model. This is crucial for improving robustness and out-of-distribution generalization of the model in the real world. We learn these invariant representations by maximizing agreement between different style interventions of the same observation via a contrastive loss in the latent feature space. The world model optimizes feature learning and controller separately to improve the sample efficiency and simplify controller learning. Motivated by the fact that most of the complexity of model-based RL approaches resides in the world model (i.e. the feature extraction and the dynamics model), we hypothesize that an additional supervisory signal from an auxiliary task helps to learn a better state representation. In this work, we used InfoNCE [34] style loss to learn invariant features. Hence, our encoder takes RGB observations (o_t) as inputs and decodes rewards (r_t) as well as depth (o_t^d) as auxiliary outputs during the training phase. The invariant feature learning is enforced by contrastive loss ($\mathcal{L}_t^q$). The proposed invariant features learning technique has the following three sub-modules:

- A *style intervention* module that utilizes data augmentation. We use spatial jitter, Gaussian blur, color jitter, grayscale and cutout data augmentation techniques for style intervention. Spatial jitter is implemented by first padding and then performing random crop. Given any observation o_t, our style intervention module randomly transforms it into two correlated views of the same observations, used for contrastive learning. All the hyperparameters are provided in the appendix.
- We use an *encoder* network that extracts representations from augmented observations, which is $\tilde{s_t} = encoder(o_t)$. The encoder is optimized using contrastive loss (Equation 1), which by construction makes the encoder invariant to the augmentations. We additionally employ an EMA regularization [16] on the encoder for stabilizing training. The hyperparameters of the encoder are provided in the appendix.
- *Contrastive loss* is defined for a contrastive prediction task, which can be explained as a differentiable dictionary lookup task. Given a query observation q and a set of keys $K = \{k_0, k_1, ...k_{2B}\}$ of length $2B$, with known positive $\{k_+\}$ and negative $\{k_-\}$ keys, the aim of contrastive loss is to learn a representation in which positive sample pairs stay close to each other while negative ones are far apart. In contrastive learning q, K, k_+ and k_- are also known as *anchors*, *targets*, *positive* and *negative* samples. We use bilinear products for *projection head* W and InfoNCE loss for contrastive learning [34], which enforces the desired similarity in the embedding space:

$$\mathcal{L}_t^q = -\log \frac{\exp(q^T W k_+)}{\exp(q^T W k_+) + \sum_{i=0}^{2(B-1)} \exp(q^T W k_i)} \quad (1)$$

3.1.2 Intervention-invariant Regularizer

Model-based RL approaches such as DreamerV2, separate the training of the world model and controller. This makes model-based RL more sample efficient as compared to model-free RL, since future predictions from latent dynamics (rollouts) can be used to learn RL policy. However, the policy learning is disconnected from world-model learning. Therefore, the supervisory signal from the policy does not optimize the encoder. While the image reconstruction loss in DreamerV2 caters to this issue for in-distribution learning, it is not enough to overcome the out-of-distribution generalization. Contrastive learning further elevates the issue, since it trains the encoder under different data augmentations while the world model still aims to predict the unaugmented image using latent dynamics. This leads to the collapse of encoder features due to a lack of correct supervisory signals. Therefore the brute-force combination of world-model and contrastive loss fails (as can be seen in Table 1, ReCoRe vs. ReCoRe-D). To this end, we propose regularization in the form of an intervention invariant auxiliary task to explicitly enforce invariance, which is robust against changes to the nuisance variables. For navigation tasks, we choose depth reconstruction $p_\theta(\widehat{d}_t|s_t, h_t)$ to verify our proposal since geometrical information remains invariant to appearance variations. The regularization task enforces the latent state to predict consistent pixel-wise depth under image augmentations by minimizing the negative of log-likelihood loss ($\mathcal{L}_t^d = -\ln q(o_t^d|s_t, h_t)$). We also experiment with image denoising and semantic segmentation on DeepMind control suite [33], where depth is not available. However, denoising is not truly invariant to RGB observation hence the performance gain on DeepMind control suite is limited but dense semantic segmentation improves the results. Other examples of intervention invariant auxiliary tasks for our augmentations are dense scene flow or sparse landmarks detection. Furthermore, our idea paves the foundation and generalizes easily to other interventions for future tasks. The key challenge is to design different interventions and intervention-invariant auxiliary tasks.

3.1.3 Learning Latent Dynamics

We leverage the DreamerV2 approach to train our world model. Similar to [13], the latent dynamics are modeled as a recurrent state space model which relies on recurrent neural network f_θ that utilizes its deterministic hidden state h_t to predict the prior stochastic state $\widehat{s}_t$. This enables efficient latent imagination for planning [10, 12]. Thus, dynamics models and representation learning modules are tightly integrated as world model and have the following components:

$$\begin{aligned}
&\text{Recurrent model:} && h_t = f_\theta(h_{t-1}, s_{t-1}, a_{t-1})\\
&\text{Representation model:} && p_\theta(s_t|h_t, \tilde{s_t})\\
&\text{Reward prediction model:} && q_\theta(\widehat{r}_t|s_t, h_t)\\
&\text{Latent dynamics model:} && q_\theta(\widehat{s}_t|h_t).
\end{aligned} \tag{2}$$

The world model is represented by neural networks and θ represents their combined parameters. It is optimized jointly with the invariant representation learning module and intervention-invariant regularizer, by minimizing,

$$\mathcal{L}_{WM} = E_p\left(\sum_t \left(\mathcal{L}_t^q + \mathcal{L}_t^d + \mathcal{L}_t^r + \beta\mathcal{L}_t^{KL}\right)\right) \tag{3}$$

where, $\mathcal{L}_t^d = -\ln q(o_t^d|s_t)$, $\mathcal{L}_t^r = -\ln q(r_t|s_t)$ and $\mathcal{L}_t^{KL} = KL(p(s_t|s_{t-1}, a_{t-1}, \tilde{s_t})||q(s_t|s_{t-1}, a_{t-1}))$.

3.1.4 Learning Controller

The objective of the controller is to optimize the expected rewards of the action, which is optimized using an actor critic approach. The actor critic approach considers the rewards beyond the horizon. Inspired by world model [10] and Dreamer [12], we learn an action model and a value model in the imagined latent space of the world model. The action model implements a policy that aims to predict future actions that maximizes the total expected rewards in the imagined environment. Given H as the imagination horizon length and γ as the discount factor for the future rewards, the action and value model are defined as follows:

$$\begin{aligned}
&\text{Action model:} && q_\phi(\widehat{a}_t|\widehat{s}_t)\\
&\text{Value model:} && E_{q(\cdot|\widehat{s}_\tau)}\textstyle\sum_{\tau=t}^{t+H}\gamma^{\tau-t}\widehat{r}_\tau.
\end{aligned} \tag{4}$$

3.2. Implementation Details

The proposed ReCoRe expands on the publicly available code base of DreamerV2 [13]. Following MoCo [16] and BYOL [8] we have used the moving average version of the query encoder to encode the keys K with a momentum value of 0.999. The contrastive loss is jointly optimized with the world model using Adam [17]. We have used five layers encoder with a starting number of feature maps equal to 32, then doubled in every consecutive layer. To encode the task observations we used two dense layers of size 32 with ELU activations [5]. The features from RGB image observation and task observation are concatenated before sending to the representation model. Replay buffer capacity is $3e^5$ for both 100k and 500k steps experiments. We update the model parameters on every fifth interactive step. Further, all architectural details and hyperparameters are provided in the appendix. The training time of ReCoRe is around 3 days on a workstation with two Nvidia GeForce RTX 3090 for 500k steps, which is twice higher than the closest state-of-the-art model-based RL model DreamerV2 [13].

4. Experiments

In our evaluation, we test ReCoRe on 3 datasets and compare it to the state of the art in model-based and model-free RL. Specifically, we use the *PointGoal* navigation task from the iGibson 1.0 environment [30] to evaluate out-of-distribution (OoD) generalization, and include the Gibson dataset [37] to test sim-to-real performance. Here we follow common practice as we compare performance based on *Success Rate* (SR) and *Success weighted by (normalized inverse) Path Length* (SPL) at 100k and 500k environment steps, which tests sample efficiency [13, 19, 20, 38]. We

		Ihlen_0_int		Ihlen_1_int		Rs_int		**Env Avg**	
Models	Steps	SR	SPL	SR	SPL	SR	SPL	SR	SPL
RAD	100k	0.6	0.01	0.1	0.00	0.8	0.01	0.5	0.01
CURL	100k	8.0	0.07	0.6	0.01	5.4	0.05	4.7	0.04
MWM	100k	1.6	0.01	0.5	0.00	2.9	0.02	1.7	0.01
DreamerV2	100k	1.8	0.01	0.6	0.00	1.7	0.01	1.3	0.01
DreamerV2 + DA	100k	7.3	0.05	1.6	0.01	7.7	0.05	5.5	0.04
DV2+G DINO	100k	**48.9**	**0.45**	**17.0**	**0.14**	45.4	0.38	**37.1**	**0.33**
ReCoRe	100k	44.8	0.38	12.2	0.09	**50.9**	**0.41**	36.0	0.29
ReCoRe - CL	100k	1.0	0.01	0.5	0.00	2.3	0.01	1.3	0.01
ReCoRe - CL + DA	100k	5.2	0.03	1.3	0.01	8.3	0.05	4.9	0.03
ReCoRe - D	100k	0.1	0.00	0.0	0.00	0.0	0.00	0.0	0.00
ReCoRe - D + I	100k	15.4	0.12	4.6	0.04	17.1	0.12	12.4	0.09
RAD	500k	48.8	0.44	11.6	0.11	48.5	0.44	36.3	0.33
CURL	500k	40.8	0.37	11.4	0.10	41.9	0.36	31.4	0.28
MWM	500k	2.6	0.02	0.7	0.00	4.2	0.02	2.5	0.01
DreamerV2	500k	1.3	0.01	0.8	0.01	2.3	0.02	1.5	0.01
DreamerV2 + DA	500k	14.7	0.10	3.9	0.03	20.5	0.13	13.0	0.08
DV2+G DINO	500k	60.9	0.58	23.2	0.19	65.8	0.58	50.0	0.45
ReCoRe	500k	**75.3**	**0.65**	**26.5**	**0.20**	**77.3**	**0.65**	**59.7**	**0.50**
ReCoRe - CL	500k	5.5	0.04	1.4	0.01	8.0	0.05	5.0	0.03
ReCoRe - CL + DA	500k	27.1	0.19	7.8	0.05	31.5	0.22	22.1	0.16
ReCoRe - D	500k	0.9	0.01	0.2	0.00	1.2	0.01	0.8	0.01
ReCoRe - D + I	500k	28.6	0.21	6.6	0.05	22.3	0.15	19.2	0.13

Table 1. Out-of-distribution generalization results on PointGoal navigation task from iGibson 1.0 dataset. *Success rate* (SR) and *Success weighted by (normalized inverse) Path Length* (SPL) are shown. We have trained on five scenes, and tested on held-out three scenes and visual textures. Our proposed world model with ReCoRe outperforms state-of-the-art RL models RAD, CURL, DreamerV2 and MWM and DV2+G DINO on 500K interactive steps, while it is on par with DV2+G DINO on 100K steps. Even though *data augmentation* (DA) improves the DreamerV2, the proposed invariant features learning technique with *contrastive loss* (CL) and *intervention invariant auxiliary task* (D) is significantly better. ReCoRe collapses when we remove the auxiliary D, however replacing with common *RGB image* reconstruction (I) recovers the performance slightly. Where, $-$ denotes 'without' and $+$ denotes 'with'.

further report on results for the DMControl suite [33], and present an ablation study.

4.1. Baselines

We compare our approach against state-of-the-art model-free RL methods RAD [19] and CURL [20], model-based RL method DreamerV2 [13], recent transformer based Masked World Model (MWM) [29] and language guided foundation model Grounding DINO [23]. We train RAD, CURL, DreamerV2 and MWM from scratch with the respective hyperparameters proposed by the authors using the official source code provided. For Grounding DINO we freeze the visual encoder and train the world model and policy (DreamerV2) from scratch using the features from the visual encoder, which we refer to as DV2+G DINO. We believe that DV2+G DINO is the strongest baseline since it has been pretrained on a large amount of data, furthermore language guidance makes the representation highly robust. Through our experiments in Table 1,2 and 3, it can be seen that (i) RAD simply overfits to the in-distribution data, though after observing large amounts of data (500K environment steps) its performance improves over CURL. (ii) DreamerV2 lacks any data augmentation in the encoder like CURL/RAD, therefore it performs significantly worse for OoD generalization. (iii) MWM relies on masking that forces the encoder to learn missing data, therefore it performs slightly better than DreamerV2. However, masking only helps in interpolation and not in extrapolation, consequently the OoD generalization is poor. (iv) Being a pretrained model DV2+G DINO shows high sample efficiency for a smaller amount of data and is on par with our approach. However, after observing a sufficient amount of data our approach shows significantly better performance.

		Ihlen		Muleshoe		Uvalda		Noxapater		McDade		**Env Avg**	
Models	Steps	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
RAD	100k	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00
CURL	100k	5.9	0.05	3.8	0.03	5.1	0.04	5.9	0.05	12.8	0.11	6.7	0.06
DV2 + G DINO	100k	38.0	**0.34**	33.5	0.29	39.7	**0.35**	37.7	**0.33**	**55.1**	**0.51**	40.8	**0.36**
ReCoRe	100k	**39.1**	0.32	**38.6**	**0.31**	**40.9**	0.32	**41.1**	**0.33**	48.2	0.39	**41.6**	0.33
RAD	500k	26.4	0.23	27.5	0.24	28.5	0.25	28.6	0.25	40.0	0.34	30.2	0.26
CURL	500k	36.8	0.33	29.3	0.27	33.7	0.30	35.2	0.32	53.8	0.50	36.7	0.33
DV2 + G DINO	500k	60.2	0.56	58.0	0.53	58.9	0.54	58.5	0.54	66.2	**0.64**	60.3	0.56
ReCoRe	500k	**74.3**	**0.64**	**72.9**	**0.62**	**73.5**	**0.61**	**72.0**	**0.61**	**67.0**	0.57	**71.9**	**0.61**

Table 2. iGibson-to-Gibson dataset: sim-to-real perception transfer results on navigation task. We choose *success rate* (SR) and *Success weighted by (normalized inverse) Path Length* (SPL) for evaluation of the models. We have trained the models on the artist created textures of iGibson, and tested on five held-out scenes from the Gibson Dataset, which are 3D scan of the real scenes. Our proposed ReCoRe outperforms state-of-the-art RL models RAD, CURL and even model-based RL combined with language guided foundation models (DV2+G DINO) on 100k and 500k interactive steps.

4.2. Out-of-Distribution Generalization

We have tested our proposed ReCoRe on random *PointGoal* navigation tasks of the iGibson 1.0 environment [30] for OoD generalization. The iGibson dataset contains 15 floor scenes with 108 rooms. The scenes are replicas of real-world homes with artist designed textures and materials. RGB, depth and task related observation (*i.e.* goal location, current location, and linear and angular velocities of the robot) are used. We emphasize, depth is only used during training. Actions include rotation in radians and forward distance in meters for the TurtleBot. We split iGibson for OoD generalization, as we chose five scenes for training and tested on the held-out three scenes. We also held out visual textures for all object classes. The details are provided in the appendix.

We report the average SR and SPL on the held-out data in Table 1 after three training runs with random seeds. Our proposed ReCoRe outperforms state-of-the-art model-based RL method DreamerV2 and model-free methods RAD and CURL on 100k and 500k interactive steps. Since DreamerV2 is natively not trained with *data augmentation* (DA), we include DreamerV2 + DA in our evaluation to show a fair comparison as all other methods contain DA inherently. We furthermore include transformer backed MWM and language guided foundation model DV2 + G DINO. Nevertheless, the proposed ReCoRe still outperforms the competitors through the invariant features learning technique with *contrastive loss* (CL) and *regularization in the form of invariant auxiliary task* (D). Further, our results show that data augmentation can also improve model-based RL, which was previously shown only for model-free RL methods [19].

4.3. Sim-to-Real Transfer

We use the Gibson dataset [37] for sim-to-real transfer experiments of the perception module, i.e. representation learning module of the world model; however please note that the robot controller is still a part of the simulator. Gibson scenes are created by 3D scanning real scenes, and a neural network is used to fill pathological geometric and occlusion errors. We have trained all models on the artist created textures of iGibson and tested on five scenes from Gibson. Table 2 shows the results. Our proposed ReCoRe significantly outperforms RAD, CURL and is slightly better compared to DV2+G DINO on 100k interactive steps, while on 500k steps ReCoRe significantly surpasses all the baselines. This shows that ReCoRe learns more stable features and is better suited for sim-to-real transfer.

4.4. Generalizable Auxiliary Tasks

The results for the DMControl suite [33] experiments are shown in Table 3. We have used image denoising (I) and semantic segmentation (S) as the intervention invariant auxiliary tasks in these experiments. ReCoRe achieved competitive results, and the key findings are: i) even though depth reconstruction is an ideal task to enforce the invariant features learning on ReCoRe explicitly, the competitive results, even without depth reconstruction show the wider applicability of the proposed model; ii) ReCoRe outperform CURL on 8 out of 12 experiments, the closest state-of-the-art RL method with contrastive learning; iii) better invariant regularizer yields better results i.e. ReCoRe produces better results with segmentation than denoising as an auxiliary task. Hence, we can conclude that ReCoRe is also competitive with end-to-end deep RL techniques even when training and evaluation environments come from similar distri-

100k Steps Total Rewards	ReCoRe+I	ReCoRe+S	CURL	Dreamer	RAD	SAC+AE	Pixel SAC
Finger, spin	486±191	474±53	**767±56**	341±70	856±73	740±64	179±66
Cartpole, swingup	472±67	449±121	**582±146**	326±27	828±27	311±11	419±40
Reacher, easy	327±98	**982±9**	538±233	314±155	826±219	274±14	145±30
Cheetah, run	321±78	**400±56**	299 ±48	235± 137	447±88	267±24	197±15
Walker, walk	654±100	**739±133**	403±24	277±12	504±191	394±22	42±12
Ball in cup, catch	830±118	**859±287**	769±43	246±174	840±179	391±82	312±63
500K Steps Total Rewards							
Finger, spin	471±173	591±181	**926±45**	796±183	947±101	884±128	179±166
Cartpole, swingup	675±64	777±64	**841±45**	762±27	863±9	735±63	419±40
Reacher, easy	891±72	**955±38**	929±44	793±164	955±71	627±58	145±30
Cheetah, run	633±70	**731±51**	518±28	570±253	728±71	550±34	197±15
Walker, walk	**965±4**	960±2	902±43	897±49	918±16	847±48	42±12
Ball in cup, catch	950±20	**984±5**	959±27	879± 87	974±12	794± 58	312± 63

Table 3. Experiment results on DMControl. Results are reported as averages across 10 seeds. ReCoRe with segmentation (S) as a regularizer achieves state-of-the-art performance over competitors in the literature on 8 out of 12 experiments, where models are targeted for other (new) downstream tasks as well. Additionally, we report RAD [19], SAC+AE [39], and Pixel SAC [11] which do not consider additional downstream tasks. However, as explained in the literature review if an agent does not need to consider the downstream tasks then optimization of the additional constraints as in ReCoRe and CURL is not necessary, and thus performance is expected to improve (this is also noted by CURL [20]). Thus ReCoRe, CURL and Dreamer form the main comparison, as these methods optimize the features learning and the controller separately.

butions (and no OoD generalization is necessary).

Additionally we include other baseline models RAD [19], SAC+AE [39], and Pixel SAC [11]. However, as explained in the literature review, if an agent does not need to consider the downstream tasks then optimization of the additional constraints (as in ReCoRe and CURL) hinders the performance (also noted by CURL [20]). So ReCoRe, CURL and Dreamer are the main competitors, as they optimize the feature learning and the controller separately. We have noticed that hyperparameter optimization for individual DMC tasks yields better results. However, we used the same set of parameters for all the tasks with ReCoRe.

4.5. Ablation Study

The contribution of contrastive learning and regularization in the form of intervention invariant auxiliary tasks are also shown in Table 1. The standard formulation of contrastive learning does not use reconstruction loss [3, 8, 16, 20]. Since model-based RL does not optimize the representation learning and controller jointly, contrastive loss collapses. Hence, to validate our proposal of intervention invariant depth reconstruction as a regularizer, we have done experiments without depth reconstruction (ReCoRe - D). We can see in Table 1 that in a reasonably complex pixel-based control task, ReCoRe is not able to learn meaningful control without the reconstruction task. Further, reconstructing RGB image (I) instead of depth (D), i.e. ReCoRe - D + I, slightly improves the results over ReCoRe - D, but is still approximately three times worse than the proposed ReCoRe.

ReCoRe without contrastive loss (ReCoRe - CL) is unable to learn meaningful control. However, data augmentation (ReCoRe - CL + DA) slightly improves these results, but is still significantly worse than the competitors on OoD generalization. Hence, these results confirm that our proposal of doing an intervention on RGB observation space and adding intervention invariant reconstruction of depth as a regularizer is a crucial necessity for facilitating the proposed world model with invariant features.

5. Conclusion

We proposed a method to learn a *World Model with invariant features*, ReCoRe. These invariant features are learned by minimizing contrastive loss between content invariance interventions of the observation. Hence, we proposed an auxiliary task as a regularizer, which is invariant to the proposed data augmentation techniques. ReCoRe significantly outperforms the state-of-the-art models on OoD generalization, sim-to-real transfer and sample efficiency measures. As such, ReCoRe is a new state of the art in model-based RL for sample efficiency in OoD generalization. Finally, we note that our framework can be applied to other tasks and the design of interventions and invariant auxiliary losses will become an interesting research problem.

References

[1] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. In *Advances in Neural Information Processing Systems*, 2019. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2, 8

[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. 2

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 5

[6] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML)*, pages 465–472, 2011. 1

[7] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2014. 2

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. 5, 8

[9] Nico Gürtler, Sebastian Blaes, Pavel Kolev, Felix Widmaier, Manuel Wüthrich, Stefan Bauer, Bernhard Schölkopf, and Georg Martius. Benchmarking offline reinforcement learning on real-robot hardware. *arXiv preprint arXiv:2307.15690*, 2023. 1

[10] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018. 1, 2, 3, 5

[11] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. 8

[12] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. 4, 5

[13] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. 1, 3, 5, 6

[14] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3

[15] Jessica B Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29, 2019. 3

[16] 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 (CVPR)*, 2020. 2, 4, 5, 8

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014. 2

[19] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems*, 2020. 4, 5, 6, 7, 8

[20] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119*, 2020. arXiv:2004.04136. 2, 4, 5, 6, 8

[21] 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. 1

[22] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 2

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 6

[24] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. 2, 3

[25] 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. 1

[26] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. 3

[27] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[28] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 2

[29] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Proceedings of The 6th Conference on Robot Learning*, pages 1332–1344. PMLR, 2023. 3, 6

[30] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, Micael Tchapmi, Kent Vainio, Josiah Wong, Li Fei-Fei, and Silvio Savarese. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2, 5, 7

[31] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018. 1

[32] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990. 3

[33] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6, 2020. 2, 5, 6, 7

[34] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2, 4

[35] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 1

[36] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, pages 2226–2240. PMLR, 2023. 3

[37] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2, 5, 7

[38] 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. 4, 5

[39] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 8

[40] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2