VISUAL CONTROL WITH VARIATIONAL CONTRASTIVE DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

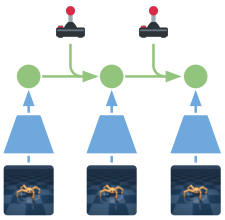
To achieve goals in complex environments with image inputs, intelligent agents must generalize to unseen situations. World models provide a way to summarize past experience to facilitate generalization by letting the agent imagine counterfactual scenarios. Recent advances in deep learning have enabled learning world models directly from images through pixel reconstruction. However, in visually complex environments, this approach requires high model capacity to succeed. We present Variational Contrastive Dynamics (VCD), a latent world model learned purely by contrasting the embeddings of encoded images. Learning behaviors in the latent space of VCD closes the gap between approaches based on contrastive learning and pixel reconstruction reconstruction, as demonstrated on both discrete Atari games and continuous visual control tasks.

1 INTRODUCTION

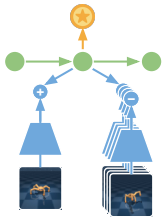
To achieve goals in complex environments with high-dimensional inputs, agents must generalize to unseen situations. One approach to this challenge is to learn a world model that represents the causal structure of the environment (Sutton, 1991; Rezende et al., 2020). Trained from past experience, the world model can then be used to make counterfactual predictions about the future to enable planning algorithms. Especially when inputs are high-dimensional images, the model can predict ahead in a compact feature space where the dynamics may become easier to learn (Watter et al., 2015). Such latent dynamics models have the potential to facilitate long-term prediction and can greatly reduce the computational and memory requirements during planning (Hafner et al., 2019b; Zhang et al., 2019).

Ideally, the latent states of a world model contain the information necessary for the agent to achieve its goals. In principle, this can be learned simply by predicting rewards (Gelada et al., 2019; Schrittwieser et al., 2019). However, this may provide an insufficient learning signal for learning world models from pixels, motivating the use of auxiliary learning objectives (Jaderberg et al., 2016; Gregor et al., 2019). A common choice is to reconstruct pixels from the latent states, which can improve sample efficiency in reinforcement learning (RL) (Kendall et al., 2019; Gregor et al., 2019). Unfortunately, pixel reconstruction can be challenging in visually complex environments and fail if the model has insufficient capacity. An alternative is contrastive learning, that learns by distinguishing images from different sequences (Gutmann and Hyvärinen, 2010). Contrastive learning focuses on what changes the most in the input and can reduce the required model capacity.

We present Variational Contrastive Dynamics (VCD), a scalable latent dynamics model that learns by contrasting embeddings. VCD builds upon recent successes in control with learned latent dynamics (Hafner et al., 2019a). By deconstructing the evidence lower bound, we identify a natural way to extend it to contrastive learning while preserving the simplicity regularizer that can be critical for learning long-term dependencies. Empirically, VCD is shown to close the gap to more expensive reconstruction-based approaches on 40 discrete and continuous control tasks.



(a) Representation



(b) Learning

Figure 1: Overview of VCD. (a) Given the image embeddings and actions, we predict compact latent states. (b) Rewards and feature vectors are predicted from each state. The contrastive loss pulls the feature vector toward the current image embedding and pushes it away from others in the batch.

2 VARIATIONAL CONTRASTIVE DYNAMICS

In this section we develop VCD in terms of its individual components as well as the objectives used.

Latent Dynamics The following three components define a latent dynamics model with model states s_t. The representation model allows us to condition the states on embeddings h_t of observed images o_t. The transition lets us predict into the future without knowing the corresponding images, given a potential action sequence. The reward predictor evaluates a state sequence,

$$\begin{aligned} &\text{Image encoder:} && h_t = e_\phi(o_t) \\ &\text{Representation model:} && p_\phi(s_t|s_{t-1}, a_{t-1}, h_t) \\ &\text{Transition model:} && q_\phi(s_t|s_{t-1}, a_{t-1}) \\ &\text{Reward model:} && q_\phi(r_t|s_t). \end{aligned} \tag{1}$$

Notationally, as in (Hafner et al., 2019a), p is used for the distributions of sampled random variables and q is used for distributions that approximate them. The model in Equation 1 can be interpreted as a non-linear Kalman filter, latent state-space model, or continuous-valued hidden Markov model.

Evidence Lower Bound When learning latent dynamics models from high-dimensional image inputs, a key design decision is the representation learning objective (Gregor et al., 2019). Given a dataset $X = x$ of images, actions, and rewards, we want to infer the corresponding model states S and learn the model parameters W.

Given a model q_ϕ, the rational belief is the posterior $q_\phi(s|x)$. However, this requires inverting the latent dynamics, which is intractable for most flexible models. Instead, we can maximize the evidence lower bound (ELBO) to find a tractable distribution $p_\phi(s|x)$ that approximates the posterior and to learn the parameters,

$$\max_\phi \mathbb{E}_{p_\phi(s|x)}[\ln q_\phi(x|s)] - \mathrm{D}_{\mathrm{KL}}[p_\phi(s|x) \parallel q_\phi(s)]. \tag{2}$$

The representation model $p_\phi(s_t|s_{t-1}, a_{t-1}, o_t)$ predicts the model states from the data using representation parameters ϕ_r. For tractability, we maintain a point estimate $\delta(w - \phi_m)$ of the model parameters. Under the factorization from Equation 1, the ELBO factors into a sum $\sum_{t=1}^{T} \left(\mathcal{J}_{\mathrm{R}}^t - \mathcal{J}_{\mathrm{D}}^t\right)$ over time steps, where

$$\begin{aligned} \mathcal{J}_{\mathrm{R}}^t &\triangleq \mathrm{E}_{p_\phi(s_t|s_{t-1}, a_{t-1}, o_t)}[\ln q_\phi(x_t|s_t)] \\ \mathcal{J}_{\mathrm{D}}^t &\triangleq \mathrm{E}_{p_\phi(s_{t-1}|s_{t-2}, a_{t-2}, o_{t-1}) p_\phi(s_{t-2}|\dots)}\left[\mathrm{D}_{\mathrm{KL}}[p_\phi(s_t|s_{t-1}, x_t) \parallel \ln q_\phi(s_t|s_{t-1})]\right]. \end{aligned} \tag{3}$$

The reconstruction term $\mathcal{J}_{\mathrm{R}}^t$ encourages mutual information between model states and images by reconstructing images from model states. The divergence term $\mathcal{J}_{\mathrm{D}}^t$ encourages simplicity by penalizing the divergence from the representation or approximate posterior from the temporal prior.

Contrastive Learning The ELBO is understood as maximizing the mutual information between the model state S_t at each time step and the corresponding image O_t, subject to a complexity regularizer. Besides reconstruction, contrastive estimation (Gutmann and Hyvärinen, 2010) offers an alternative for maximizing the mutual information.

Reconstruction (Equation 3) estimates the mutual information by predicting the image from the latent state. Subtracting the constant log-likelihood of the data again makes this explicit. Contrastive estimation instead estimates the mutual information by predicting the latent state from the image,

$$\begin{aligned} &\text{Mutual information:} && \mathrm{E}[\ln p_\phi(x_t, s_t) - \ln p(x_t) - \ln p_\phi(s_t)] \\ &\text{Recon bound:} && \mathrm{E}[\ln q_\phi(x_t \mid s_t) - \ln p(x_t)] \\ &\text{Contrastive bound:} && \mathrm{E}[\ln q_\phi(s_t \mid x_t) - \ln p_\phi(s_t)]. \end{aligned} \tag{4}$$

Where the marginal in image space was a constant, contrastive estimation uses the marginal latent that depends on our representation model. The marginal is approximated using conditionals averaged across the current minibatch. This estimator is known as InfoNCE (Oord et al., 2018) and bounds

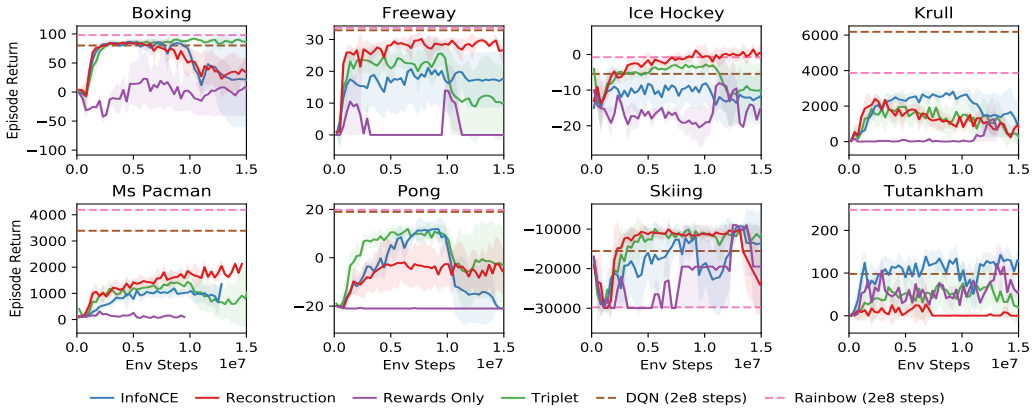


Figure 2: Performance of contrastive objectives and baselines on selected Atari environments. The contrastive objectives display performance over or on-par with reconstruction.

mutual information (Poole et al., 2019),

$$\begin{aligned}\mathcal{J}_{\mathrm{C}}^{t} &\triangleq \mathrm{E}_{p_\phi(s_t \mid s_{t-1}, a_{t-1}, o_t)}\Big[\ln q_\phi(s_t \mid x_t) - \ln \sum_{x'} q_\phi(s_t \mid x')\Big] \\ \mathcal{J}_{\mathrm{D}}^{t} &\triangleq \mathrm{E}_{p_\phi(s_{t-1} \mid s_{t-2}, a_{t-2}, o_{t-1}) p_\phi(s_{t-2} \mid \ldots)}\Big[\mathrm{D}_{\mathrm{KL}}\big[p_\phi(s_t \mid s_{t-1}, x_t) \,\big\|\, \ln q_\phi(s_t \mid s_{t-1})\big]\Big].\end{aligned} \tag{5}$$

The discriminator $\ln q_\phi(s_t \mid x_t)$ is implemented as a neural network that uses the encoder features $h_t = e_\phi(x_t)$. The first term, $\ln q_\phi(s_t \mid x_t)$, is computed for positive pair where s_t and x_t belong to the same sequence and time step. The second term, $\ln \sum_{x'} q_\phi(s_t \mid x')$, is computed from negative pairs where s_t is the current state and x' are images in the current mini-batch. Maximizing a discriminator for the positive pair and minimizing it for all negative pairs is illustrated in Figure 1. The training objective for VCD with InfoNCE is,

$$L_{InfoNCE} = \sum_{t=1}^{T} \big(\mathcal{J}_{\mathrm{C}}^{t} + \mathcal{J}_{\mathrm{D}}^{t}\big) \tag{6}$$

Several recent representation learning methods leverage the InfoNCE bound to maximize mutual information. In contrast, we keep the KL regularizer from the ELBO objective. The resulting objective mimics the information bottleneck (Tishby et al., 2000; Alemi et al., 2016).

Triplet loss We also experiment with another type of contrastive loss: the triplet loss of Schroff et al. (2015). For a given embedding "anchor" A_t, we would like to push it closer to a "positive" embedding P_t of the same label, and farther away from a "negative" embedding N_t of a different label. Closeness is measured under some similarity metric $d(\cdot, \cdot)$, usually the Euclidean norm. Triplet loss represents this objective for a given $\{A_t, P_t, N_t\}$ tuple and a margin m as,

$$\mathcal{J}_{\mathrm{T}}^{t} = \max\big(m + d(A_t, P_t) - d(A_t, N_t), 0\big). \tag{7}$$

N_t should be chosen such that its distance to A_t is within a margin of the distance between A_t and P_t. Beyond this margin, the negative example is sufficiently different from the anchor and should not contribute to the loss. If it is within this margin, then we would like to push it further away.

The objective for VCD with triplet loss is,

$$L_{triplet} = \sum_{t=1}^{T} \big(\mathcal{J}_{\mathrm{T}}^{t} + \mathcal{J}_{\mathrm{D}}^{t}\big) \tag{8}$$

By minimizing this objective over multiple triplet pairs, we are able to learn useful representations.

3 Experiments

We evaluate VCD on a suite of Atari games (Bellemare et al., 2013) and DeepMind Control Suite (Yuval et al., 2018) tasks. We implement VCD separately with InfoNCE and triplet loss, and compare

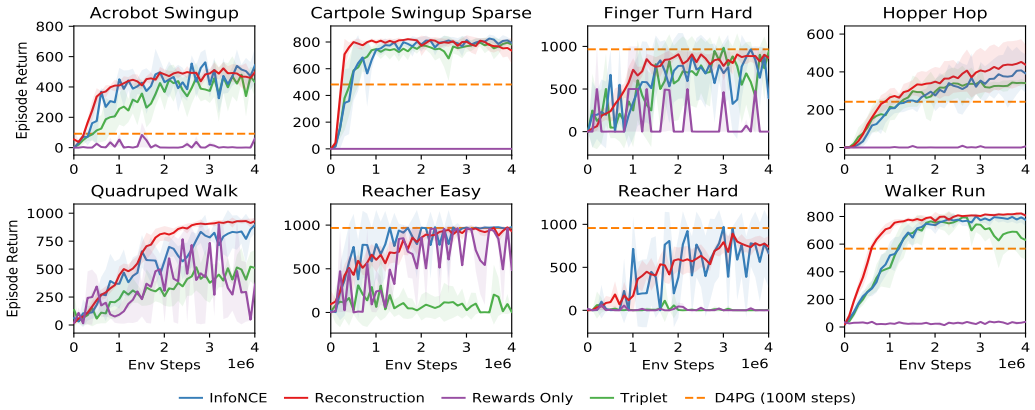


Figure 3: Performance of contrastive objectives and baselines on selected environments from the DeepMind Control Suite. InfoNCE consistently performs on-par with pixel reconstruction.

it to pixel reconstruction and reward prediction only as baselines. Each individual agent is trained on a NVIDIA V100 GPU. VCD takes 6.5 hours to train for 10^6 steps on the DeepMind Control Suite. This is a comparable result to the computational cost of learning from image reconstruction.

We use a CNN with ELU activations (Clevert et al., 2015) to encode observations (Ha and Schmidhuber, 2018) and a Recurrent State Space Model (RSSM) (Hafner et al., 2018) to model the environment dynamics. We use both learning and model hyperparameters listed in Dreamer (Hafner et al., 2019a) out of the box with no additional tuning. All reported results are averaged over 5 seeds, with the exception of the reward baseline which used 2 seeds.

- **InfoNCE** The world model is learned using VCD, where embeddings are compared along the batch dimension by an inner product between the embeddings (i.e. cosine distance).
- **Triplet** The world model is learned using VCD with triplet loss, where embeddings are compared using Euclidean norm. Negatives are chosen through semi-hard negative mining.
- **Reconstruction** The world model is learned by predicting rewards and reconstructing images using a transposed CNN, recovering the Dreamer agent (Hafner et al., 2019a).
- **Rewards Only** No auxiliary representation learning is used and the world models learns solely by predicting the scalar reward of the task under the simplicity regularizer.

We report results on 20 Atari tasks (Bellemare et al., 2013) where one or more methods achieved consistent learning, and on 20 DeepMind Control Suite (DMC) tasks (Tassa et al., 2018) where the top model-free agent D4PG (Barth-Maron et al., 2018) achieved non-zero scores. A selected subset of the learning curves can be found in Figure 2 and Figure 3, whereas complete plots of each can be found in Figure 4 and Figure 5 of the Appendix. In the Atari environments, both contrastive approaches close the gap to pixel reconstruction. Out of the 20 tasks, InfoNCE ties with it on 8, outperforms it on 4 and underperforms on 8. Triplet ties it on 9, outperforms it on 2 and underperforms on 9. The most noticeable performance gains for the contrastive approaches are on Krull, Pong, and Tutankham. These may be difficult for reconstruction to model due to the presence of small but important visual details (Pong), or due to continuously changing backgrounds (Krull and Tutankham) in the environment. The InfoNCE approach consistently matches the performance of the pixel reconstruction baseline across all DMC environments. On the other hand, the triplet approach fails on the Reacher tasks, and underperforms on both Quadruped Walk and Hopper Hop.

4 Summary

In this work we present VCD, a latent dynamics world model that learns embeddings from image inputs in a contrastive manner. State embeddings are matched with their corresponding image embedding, and contrasted with state and image embeddings from different trajectories. VCD is a lightweight model, relying on a smaller architecture than pixel reconstruction. It is also a causally correct model with respect to rewards. Despite that, it matches or outperforms pixel reconstruction on a number of control tasks from the Atari environments and the DeepMind Control Suite.

REFERENCES

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, A. Muldal, N. Heess, and T. Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. *arXiv preprint arXiv:1906.02736*, 2019.

K. Gregor, D. J. Rezende, F. Besse, Y. Wu, H. Merzic, and A. van den Oord. Shaping belief states with generative environment models for rl. In *Advances in Neural Information Processing Systems*, pages 13475–13487, 2019.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*, 2018.

D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565, 2019b.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254. IEEE, 2019.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

D. J. Rezende, I. Danihelka, G. Papamakarios, N. R. Ke, R. Jiang, T. Weber, K. Gregor, H. Merzic, F. Viola, J. Wang, J. Mitrovic, F. Besse, I. Antonoglou, L. Buesing, J. Schrittwieser, T. Hubert, and D. Silver. Causally correct partial models for reinforcement learning, 2020. URL https://openreview.net/forum?id=HyeG9yHKPr.

J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

R. S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pages 2746–2754, 2015.

T. Yuval et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

M. Zhang, S. Vikram, L. Smith, P. Abbeel, M. Johnson, and S. Levine. Solar: deep structured representations for model-based reinforcement learning. In *International Conference on Machine Learning*, 2019.

A FULL ATARI LEARNING CURVES

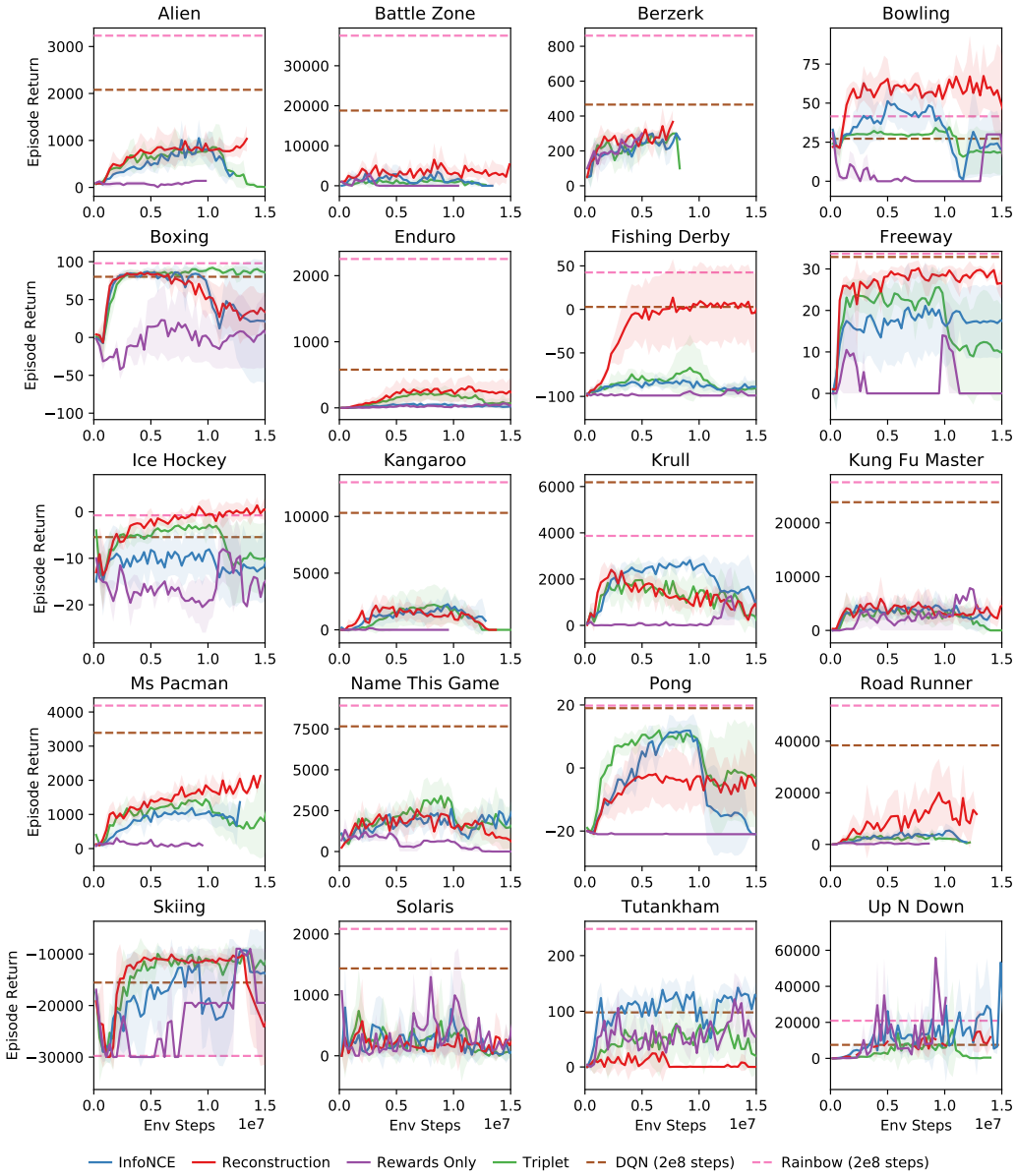


Figure 4: Atari

B FULL CONTROL SUITE LEARNING CURVES

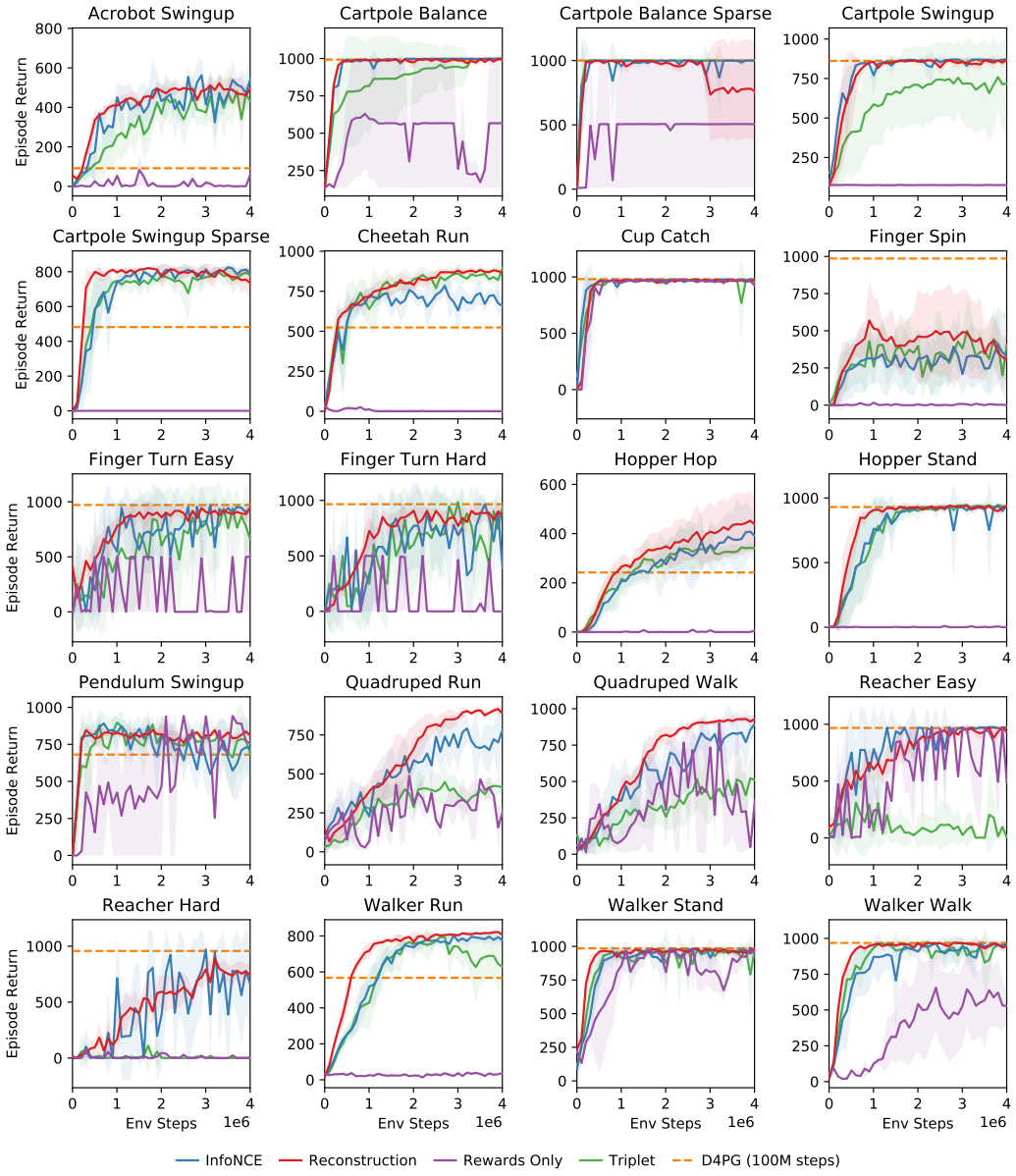


Figure 5: Control Suite