

# Self-Supervision for Reinforcement Learning

*Parsa Mahmoudieh*



Electrical Engineering and Computer Sciences  
University of California at Berkeley

Technical Report No. UCB/Eecs-2017-51

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/Eecs-2017-51.html>

May 11, 2017

Copyright © 2017, by the author(s).  
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

---

# Self-Supervision for Reinforcement Learning

by Parsa Mahmoudieh

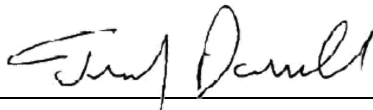
---

## Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,  
University of California at Berkeley, in partial satisfaction of the requirements for the  
degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

### Committee:



---

Professor Trevor Darrell  
Research Advisor

5/10/17

---

(Date)

\* \* \* \* \*



---

Professor Sergey Levine  
Second Reader

5/10/17

---

(Date)

### **Abstract**

Reinforcement learning optimizes policies for expected cumulative reward. Need the supervision be so narrow? Reward is delayed and sparse for many tasks, making it a difficult and impoverished signal for end-to-end optimization. To augment reward, we consider a range of self-supervised tasks that incorporate states, actions, and successors to provide auxiliary losses. These losses offer ubiquitous and instantaneous supervision for representation learning even in the absence of reward. While current results show that learning from reward alone is feasible, pure reinforcement learning methods are constrained by computational and data efficiency issues that can be remedied by auxiliary losses. Self-supervised pre-training and joint optimization improve the data efficiency and policy returns of end-to-end reinforcement learning.

# 1 Introduction

End-to-end reinforcement learning (RL) addresses representation learning at the same time as policy optimization. Of these dual pursuits, current work focuses on the reinforcement learning aspects of the problem such as stochastic optimization and exploration. Once a loss on reward is defined the representation is delegated to backpropagation without further attention to other supervisory signals. We argue that representation learning is a bottleneck in current approaches bound by reward. Our self-supervised auxiliary losses broaden the horizons of reinforcement learning agents to learn from all experience, whether rewarded or not.

To illustrate the critical role of representation learning, we show that re-training a decapitated agent, having destroyed the policy and value outputs while preserving the rest of the representation, is far faster than the initial training (Figure 1). Although the policy distribution and value function are lost, they are readily recovered given a representation from RL, even though the optimization and exploration issues remain. With the importance of representation established, we turn to self-supervision to take an ambient approach to RL attuned to reward and environment alike.

Self-supervision defines losses via surrogate annotations that are synthesized from bare, unlabeled inputs. In the context of RL, reward captures the task while self-supervision captures the environment. In this setting, every transition contributes gradients of ambient environmental signals. While reward might be delayed and sparse, the losses from self-supervision are instantaneous and ubiquitous. Augmenting RL with these auxiliary losses enriches the representation through multi-task learning and improves policy optimization.

We concentrate on auxiliary losses for state, dynamics, inverse dynamics, and reward that can be formulated in a discriminative fashion. To help RL, we transfer the representation from self-supervised pre-training with these losses. In the other direction, we inspect the contents of policy representations by examining transfer from RL to self-supervised tasks. Pre-training for Atari reaches higher returns with better data efficiency for a  $1.4\times$  speed-up on average to 95% of the best return. Joint optimization improves further still.

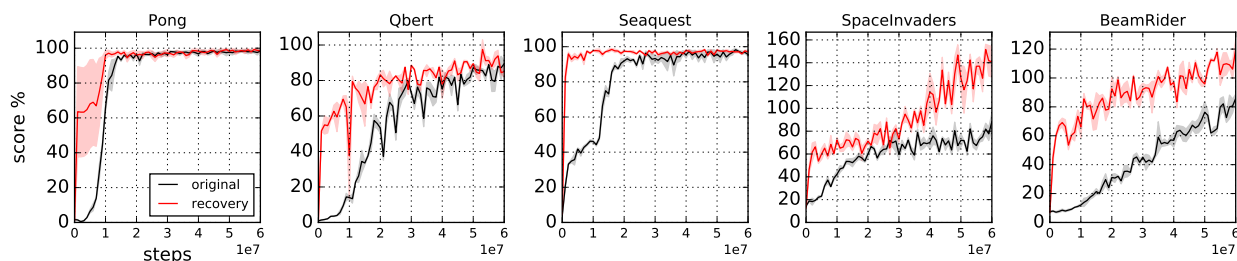


Figure 1: Current methods require many transitions to arrive at good policies, but policies are often quickly recovered from their representation. To separate reinforcement learning from representation learning, we decapitate the agent by destroying its policy and value output parameters, and then re-train end-to-end. Although the policy distribution and value estimates are obliterated, most of the parameters are preserved and the policy is swiftly recovered. The gap between the initial optimization and recovery illustrates a representation learning bottleneck.

## 2 Preliminaries

We briefly review policy gradient methods for RL and then frame self-supervised learning and relate it to supervised and unsupervised learning.

<sup>0</sup>Joint work with Evan Shelhamer, Max Argus, and Trevor Darrell <https://arxiv.org/abs/1612.07307v2>

## 2.1 Reinforcement Learning

Reinforcement learning (RL) is concerned with policy optimization on Markov decision processes (MDPs). Consider an MDP defined by the tuple  $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ , where  $\mathcal{S}$  is the set of states,  $\mathcal{A}$  is the set of actions,  $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is the transition probability distribution,  $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$  is the reward function, and  $\gamma \in (0, 1)$  is the discount. In addition let  $p_0$  be the distribution of the initial state  $s_0$ . Let  $\pi$  be a stochastic policy  $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ , and  $\pi_\theta$  be a policy parameterized by  $\theta$ .

The objective is to maximize the expected return  $\eta(\pi)$  of the policy:

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, \dots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \right], \text{ where}$$

$$s_0 \sim p_0(s_0), \quad a_t \sim \pi(a_t | s_t), \quad s_{t+1} \sim T(s_{t+1} | s_t, a_t)$$

The expected return is measured by the state-action value  $Q_\pi$ , the value  $V_\pi$ , and the advantage  $A_\pi$ :

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \dots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \right],$$

$$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \dots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \right],$$

and  $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$

where  $a_t \sim \pi(a_t | s_t)$  and  $s_{t+1} \sim T(s_{t+1} | s_t, a_t)$ .

Policy gradient methods iteratively optimize the policy return by estimating the gradient of the expected return with respect to the policy parameters

$$\nabla_\theta \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right] = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \nabla_\theta \log \pi_\theta(a_t | s_t) \right],$$

where the expectation is sampled by executing the policy in the environment. To improve optimization, in an actor-critic method the policy gradient can be scaled not by the return itself but by an estimate of the advantage Sutton & Barto (1998). In this work we augment the policy gradient with auxiliary gradients from self-supervised tasks.

## 2.2 Self-Supervision

End-to-end RL admits policy learning in lieu of policy design in much the same way that end-to-end supervised learning has seen the advance of feature learning over feature design. Supervised learning, especially as carried out for computer vision, has recently seen the rise of deeper and higher-capacity networks trained by backpropagation, reaching 100+ layers (He et al., 2016). These capacities are sustained only by massive amounts of annotation and other supervisory signals. Supervised pre-training on large-scale annotations as exemplified by ImageNet (Deng et al., 2009) currently delivers the most effective features for transfer learning to other tasks. However, a wave of renewed interest in unsupervised and “self-supervised” learning offers alternatives that we catalogue here (Doersch et al., 2015; Noroozi & Favaro, 2016; Zhang et al., 2016; Donahue et al., 2016).

To illustrate the differences, consider three kinds of learning by their objectives:

- supervised learning  $\min_\theta \mathbb{E} [L_{\text{dis}}(f_\theta(x), y)]$
- unsupervised learning  $\min_\theta \mathbb{E} [L_{\text{gen}}(f_\theta(x), x)]$
- self-supervised learning  $\min_\theta \mathbb{E} [L_{\text{dis}}(f_\theta(x), s(x))]$  with surrogate annotation function  $s(\cdot)$

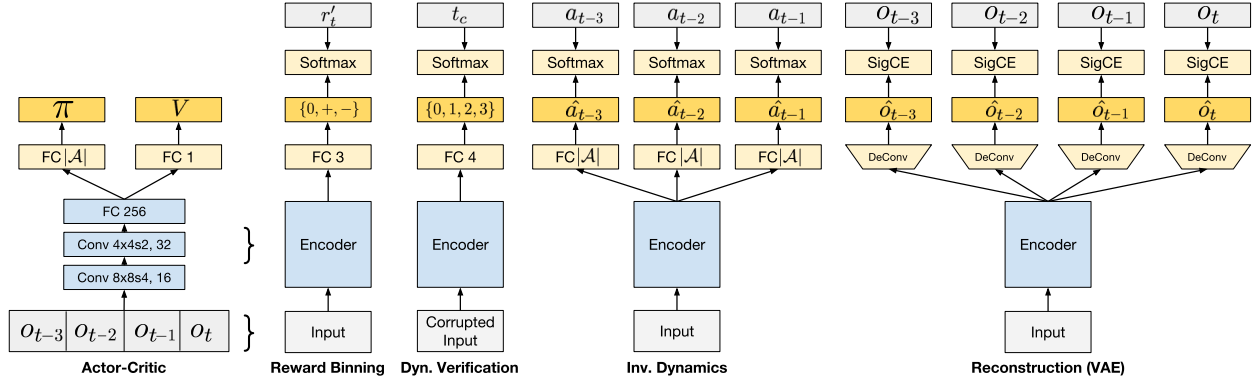


Figure 2: Architectures for reinforcement learning and self-supervision. The actor-critic architecture is based on A3C (Mnih et al., 2015) but with capacity reduced for experimental efficiency. The self-supervised architectures share the same encoder as the actor-critic for transferability. Each self-supervised task augments the architecture with its own decoder and loss.

for data  $x$ , annotations  $y$ , losses  $L$  either discriminative or generative, and parametric model  $f_\theta$ . Both unsupervised learning and self-supervised learning define losses without annotation, but unsupervised learning has historically focused on generative or reconstructive losses, while nascent self-supervised methods instead define surrogate losses and synthesize the annotations from the data. Since self-supervised and unsupervised methods can make use of unannotated data, as auxiliary losses for RL they promise to mine more from the data already available to the policy.

### 3 Self-Supervision of Policies

The state, action, reward, and successor  $(s, a, r, s')$  transition standard to RL admits many kinds of self-supervision. We explore the use of surrogate annotations that span different parts of the transitions to gauge what is informative for RL. These diverse, ambient losses mine further supervision from the same data available to existing RL methods.

Adopting self-supervision for RL raises issues of multi-task optimization and statistical dependence. Policy optimization and self-supervision may need to be reconciled to learn from both reward and auxiliary losses without interference. As for the data distribution, in the RL setting the distribution of transitions is neither i.i.d. nor stationary, so self-supervision should follow the policy distribution. We first take the simple approach of self-supervised pre-training followed by pure RL. For pre-training we only optimize auxiliary losses on the initial, random policy distribution and do not track the policy distribution. To remedy this and achieve further gains we switch to joint optimization of reinforcement learning and self-supervision.

#### 3.1 Tasks

For RL transfer, the self-supervised tasks must make use of the same transition data as RL while respecting architectural compatibility with the agent network. We first survey auxiliary losses and then define their instantiations for our chosen environment and architecture. Every self-supervised task augments a common, agent-compatible encoder with a task-specific decoder. Once pre-training is complete the decoder is discarded and the shared representation is transferred to the initial agent network. Figure 2 illustrates tasks and architectures.

**Reward** Self-supervision of reward is a natural choice to tune the representation for RL. Reward can be cast into a proxy task as instantaneous prediction by regression or binning into positive, zero, and negative classes. Our self-supervised reward task is to bin  $r_t$  into  $r'_t \in \{0, +, -\}$  with equal balancing of the classes as done independently by Jaderberg et al. (2016). This is equivalent to one-step or zero-discount value function estimation, and so may seem

redundant for value methods. However, the gradient of the instantaneous prediction task is less noisy because it is not subject to policy stochasticity or bootstrapping error. With reward, the proxy task accuracy is expected to closely mirror the degree of policy improvement.

**Dynamics and Inverse Dynamics** Surrogate annotations for these tasks capture state, action, and successor  $(s, a, s')$  relationships from transitions. Even a single transition suffices to define losses on dynamics (successors) and inverse dynamics (actions). The losses need not form a transition model, and simple proxies can suffice to help tune the representation. The difficulty of temporal self-supervision can be adjusted through the span and stride of time steps.

Dynamics can be cast into a verification task by recognizing whether state-successor  $(s, s')$  pairs are drawn from the environment or not. This can be made action conditional by extending the data to  $(s, a, s')$  and solving the same classification task. Our self-supervised dynamics verification task is to identify the corrupted observation  $o_{t_c}$  in a history from  $t_0$  to  $t_k$ , where  $o_{t_c}$  is corrupted by swapping it with  $o_{t'}$  for  $t' \notin \{t_0, \dots, t_k\}$ . We synthesize negatives by transplanting successors from other, nearby time steps. While the transition function is not necessarily one-to-one, and the synthetic negatives are noisy, in expectation these surrogate annotations will match the transition statistics.

Inverse dynamics, mapping  $\mathcal{S} \times \mathcal{S} \rightarrow \mathcal{A}$ , can be reduced to classification (for discrete actions) or regression (for continuous actions). Our self-supervised inverse dynamics task is to infer the intervening actions of a history of observations. When  $|\mathcal{A}| \ll |\mathcal{S}|$ , as is often the case, the self-supervision of inverse dynamics may be more statistically and computationally tractable.

**Reconstruction** Auto-encoding/AE Hinton & Salakhutdinov (2006) and variational auto-encoding/VAE Kingma & Welling (2014); Rezende et al. (2014) learn to reconstruct the input subject to a representational bottleneck. Generative adversarial networks/GANs Goodfellow et al. (2014) optimize a generator and discriminator to learn a model of the data, to which bidirectional GANs Donahue et al. (2016) add an encoder for adversarial feature learning. The surrogate annotation for reconstruction is simply the identity as the loss is a distance between the input and output. While a popular line of attack for unsupervised learning, the representations learned by reconstruction are relatively poor for transfer (Donahue et al., 2016). Nevertheless we include reconstruction for comparison with self-supervised tasks that map inputs to distinct surrogate annotations.

**Observation Cues** A number of visual signals have been identified that help learn transferable features. Visual coherence and context (Doersch et al., 2015; Noroozi & Favaro, 2016; Pathak et al., 2016) are cast into losses by discriminatively recognizing spatial relationships (as in solving a jigsaw puzzle) or generating input pixels (as in inpainting). Colorization of greyscale imagery (Zhang et al., 2016) or more generally any image-to-image mapping between modalities can be cast into pixelwise auxiliary losses. As the policy acts across transitions, and dependence spans time, it may be insufficient to self-supervise observations alone. In preliminary experiments these losses had no effect so we do not pursue them further.

In our approach the purpose of self-supervision is representation learning and not full modeling of the dynamics and reward. As illustrated by these proxy tasks, the surrogate annotations need not directly predict the transition and reward functions. The auxiliary losses are expected to give gradients and not necessarily furnish a generative model for model-based RL. While modeling could be intractable, the gradients might suffice to improve reinforcement learning.

### 3.2 Loss as Intrinsic Reward

Intrinsic rewards are intended to scaffold skill learning, aid exploration, or otherwise guide the policy to improve (Barto et al., 2004; Chentanez et al., 2004). Rewards that formalize novelty, curiosity, and competence focus on learning progress and predictive error (Schmidhuber, 1991; Oudeyer et al., 2005; Houthoofd et al., 2016). Self-supervisory losses could serve as intrinsic rewards of this kind, and simultaneously guide the policy while tuning the representation through gradients.

Self-supervisory intrinsic rewards could lead the policy to novel and unlearned states for exploration. Following the loss could visit the transitions to still be learned, until they learned, and then move on. It may be crucial to reward learning progress, and not the absolute loss, to ensure improvement. This in effect importance samples by the auxiliary



	Pong	Qbert	Seaquest	S. Invaders	BeamRider	Breakout
VAE [ $\ell_2$ ]	1.6	1.8	2.5	1.7	2.5	1.2
BiGAN [ $\ell_2$ ]	4.5	5.7	6.1	6.6	11.8	5.8
...obs. mode	2.48	8.34	8.00	16.13	59.7	14.4
Reward [F1]	0.99	0.82	0.03	0.38	0.16	0.90
Dyn. Ver. [acc. %]	97.5	92.8	95.0	90.5	98.6	70.8
...chance	25	...	...	...	...	...
Inv. Dyn. [acc. %]	34.9	17.5	25.5	33.3	21.1	33.9
...chance	16.6	16.6	5.5	16.6	11.1	16.6

Table 1: Feasibility of the self-supervised tasks for Atari. Most tasks reach reasonable performance. Task metrics improve through training and optimization converges quickly in less than ten epochs.

losses. A baseline for this directed pre-training is self-supervision on the static data distribution of a fixed random policy. Unifying loss and reward in this way is an underexplored opportunity supplied by end-to-end RL.

## 4 Results

We show results on self-supervision for policy pre-training and joint optimization on Atari. To begin we check the feasibility of the self-supervised tasks on transitions collected from random policies. Then for each proxy task and environment we measure improvements in return and data efficiency for self-supervised policy pre-training. As a probe into policy representations, we examine decoding from fixed reinforcement learning weights to proxy tasks. Policies trained with self-supervision converge to the same or better return and do so in fewer updates.

### 4.1 Self-Supervision

Our collection of self-supervised policies for Atari are variations of the asynchronous advantage actor-critic (A3C) architecture of Mnih et al. (2016). The actor-critic network is taken as an encoder to which each task attaches its own decoder. To begin, we optimize the proxy tasks for their own sake to check their admissibility as pre-training for RL. In general, the self-supervised tasks achieve reasonable performance and converge quickly. The task metrics across several environments are reported in Table 1. Note that proxy task performance need not be perfect to yield a transferable representation, and indeed low proxy task accuracy can still deliver state-of-the-art self-supervised features Doersch et al. (2015).

Multi-task training of the reward, dynamics, and inverse dynamics tasks achieves comparable scores. It was not necessary to balance losses or otherwise tune learning for multi-task optimization of these auxiliary losses. The encoder apparently has enough capacity to jointly address these proxy tasks. A higher-capacity encoder may do better still, and the interaction of encoder capacity, self-supervision, and policy optimization could compound gains for deeper architectures in richer environments.

### 4.2 Policy Pre-training

Self-supervised pre-training followed by RL fine-tuning is the simplest approach to incorporating auxiliary losses. This simplicity controls for confounds in joint optimization such as loss weighting and learning rate schedules. Any effect of self-supervision is purely due to representation learning prior to reinforcement learning.

We compare simple initialization strategies—random initialization as well as calibrated and data-dependent initialization (Krähenbühl et al., 2016)—with our self-supervised tasks. The calibrated random initialization has little effect while data-dependent initialization variably helps and hurts. Our self-supervised tasks boost RL further and do so in more cases. These tasks include auxiliary losses that are agnostic to reward, letting learning make progress without it.

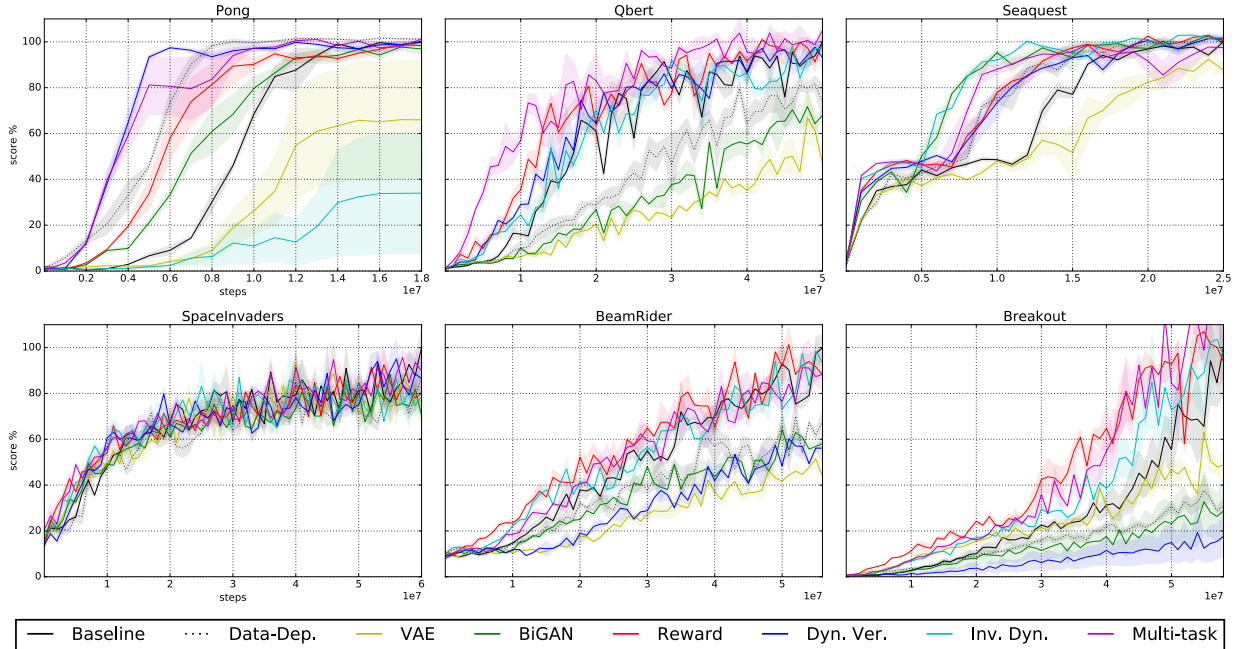


Figure 3: Optimization progress for reinforcement learning from self-supervised pre-training. Progress is reported as percentage of the best baseline return with evaluation every 1M updates. Many self-supervised tasks improve data efficiency without sacrificing return. Tasks independent of reward, such as dynamics and inverse dynamics, can nevertheless improve optimization. Improvement is strongest early in training when the pre-training and policy distributions are close. Refer to Section 3.1 for the details of the auxiliary tasks. The mean and variance of the score is calculated over three runs.

Perhaps surprisingly, self-supervision of reward is not universally the most effective pre-training. Figure 3 shows policy optimization progress with the various pre-training schemes, Figure 4 reports data efficiency, and Table 2 reports policy returns.

The immediate observation is that pre-training suffices to improve optimization. Returns at convergence are nearly equal or better than baseline and the optimization is more data efficient. The sole exception is when pre-training diverges, but this is simple to diagnose. Pre-training is most helpful early in the optimization, when the policy distribution is close to the random distribution (which is the data distribution for pre-training). In the few-shot or budgeted regimes, there is a steeper advantage to self-supervision.

Overall the self-supervised tasks surpass reconstructive tasks. Reconstruction by VAE is mostly harmful, but on the other hand BiGAN results show some improvement. However, there is no clear ordering of the individual tasks across environments, neither for policy return nor for data efficiency. Multi-task optimization of reward, dynamics, and inverse dynamics tends to improve on both fronts. When ranked by data efficiency, the median rank of the baseline across environments is 4.5 (out of 8) while multi-task pre-training ranks second. Multi-task self-supervision is a practical default.

### 4.3 Probing Policy Representations

Transfer from self-supervision to RL scaffolds the policy representation and improves optimization. However, whether transfer helps by capturing aspects of the environment or merely conditioning the weights is unclear. Furthermore, it is not obvious what is encoded by policy representations learned by pure RL.

To gather indirect evidence, we explore transfer from RL to our proxy tasks to see which can be decoded from fixed

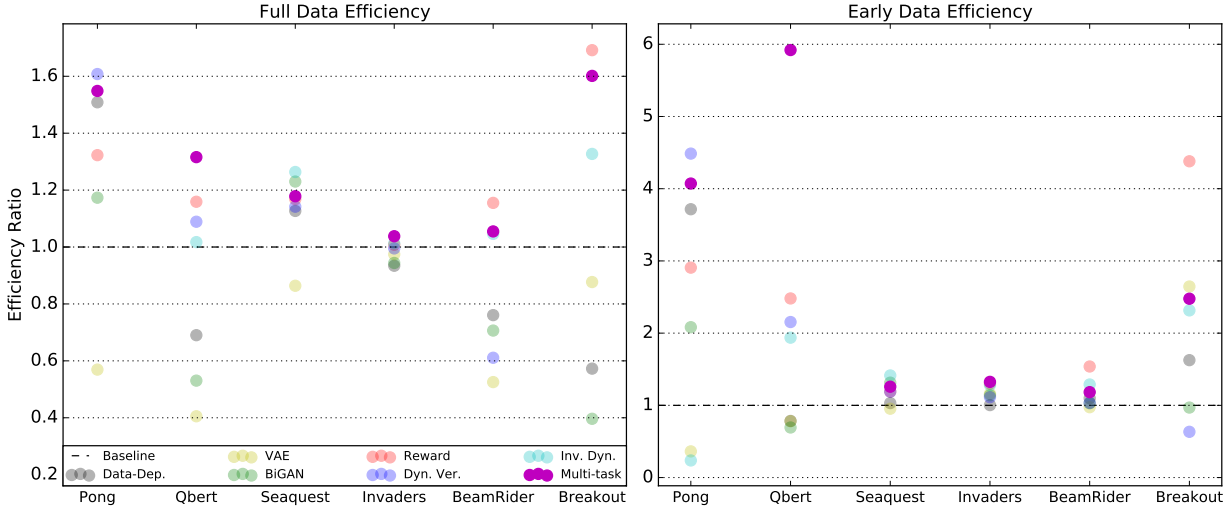


Figure 4: Data efficiency of RL with self-supervised pre-training. To measure data efficiency, we calculate the area under the score/iteration curve and report the ratio to the baseline. Multi-task self-supervision improves  $1.3\times$  on average for full optimization to 60M iterations. Focusing on early optimization, multi-task self-supervision gives  $2.7\times$  improvement for the first 10M iterations.

	Pong	Qbert	Seaquest	S. Invaders	BeamRider	Breakout
Baseline	21	18028	1756	1102	5061	367
Data-Dep.	100%	90%	100%	99%	74%	39%
VAE	100%	82%	99%	<b>107%</b>	51%	62%
BiGAN	<b>101%</b>	84%	100%	83%	61%	43%
Reward	100%	101%	100%	91%	97%	96%
Dyn. Ver.	99%	105%	101%	102%	61%	37%
Inv. Dyn.	100%	97%	101%	100%	96%	102%
Multi-task	<b>101%</b>	<b>111%</b>	<b>102%</b>	105%	<b>99%</b>	<b>110%</b>

Table 2: Returns from by self-supervised pre-training. Returns are reported as the absolute return for the baseline (pure RL from random initialization) and the return relative to the baseline for the other conditions. The returns achieved are nearly equal or better.

	Pong	Qbert	Seaquest	S. Invaders	BeamRider	Breakout
VAE	-	71%	-	72%	65%	-
Reward	99%	63%	67%	29%	25%	44%
Dyn. Ver.	91%	33%	43%	38%	42%	117%
Inv. Dyn.	56%	62%	58%	69%	62%	81%

Table 3: Analysis of proxy tasks by decoding RL representations. We measure the accuracy of learning from fixed features instead of end-to-end. The relative performance gives some indication of what is captured by pure RL features. The drops in accuracy suggest that the representation is narrowly tuned to the RL task.

parameters. For each task we affix a decoder to the feature layer from which policy and value predictions are made. The decoder is learned and evaluated on data from the policy distribution at the end of training. Table 3 reports the accuracy of learning the proxy tasks from RL parameters compared to end-to-end optimization.

Most proxy tasks suffer a significant drop in accuracy ( $>30\%$ ). The VAE even diverges for several environments. Learning and evaluating the decoder on the initial, random policy data does worse still, suggesting the representation is closely tuned to the current policy distribution. Although these same proxies can improve RL, the RL representation itself seems to be different, and perhaps narrowly tuned to the task defined by reward.

#### 4.4 Joint Policy and Auxiliary Optimization

Having shown that pre-training is effective in its own right, we turn to joint optimization to further boost the effects of self-supervision. Online, multi-task optimization guarantees that the auxiliary losses are optimized on the policy distribution. For combined supervision we simply sum the losses and gradients from reinforcement learning and self-supervision. For a comparison of joint optimization and pre-training on Pong see Figure 5.

The joint optimization results improve on the pre-training for every task. Note that inverse dynamics fails when pre-trained but improves over the baseline when trained jointly. This underscores the importance of tracking the changing

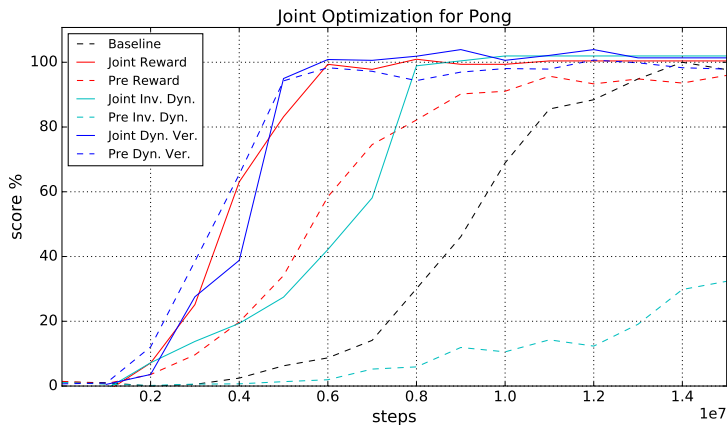


Figure 5: Joint optimization with auxiliary losses further improves over pre-training. All of the joint tasks (solid lines) have comparable or higher data efficiency than their respective pre-trained tasks (dashed lines).

policy distribution: doing so helps more than any potential interference among the RL and auxiliary losses.

## 4.5 Experimental Framework

Pre-training is carried out as straightforward supervised learning on fixed data shared by all of the tasks. The data for pre-training of each environment is collected by executing a random policy for 100,000 transitions. The transitions collected for self-supervision are pre-processed in the same format as the transitions encountered during RL. A portion of the collected pre-training episodes is held out for validation.

Transfer to reinforcement learning is carried out in the same manner across all tasks. First the self-supervised output layers are discarded and replaced by outputs for policy and value. The policy and value weights are initialized according to LeCun et al. (1998) as in Mnih et al. (2015). To control for disparities in auxiliary losses, all networks are calibrated to equalize gradients across layers by the method of Krähenbühl et al. (2016). Without calibration transfer can fail to improve over random policy performance. In rare cases, should training still fail, we fallback to transferring only the convolutional layers.

Joint optimization is carried out by summing the policy and auxiliary losses and gradients. Each auxiliary loss has its own weight selected by cross-validation and shared across environments. The policy loss is computed on-policy from rollouts while the auxiliary losses sample mini-batches from a small replay memory (<10,000 transitions).

For architecture we adapt the actor-critic network of Mnih et al. (2016) but reduce its capacity to that of the original DQN (Mnih et al., 2013) for computational efficiency. For optimization we select the state-of-the-art asynchronous advantage actor-critic (A3C) method (Mnih et al., 2016) and configure it with comparable hyperparameters. For the environment we adhere to the specification from DeepMind (Mnih et al., 2015) by our own re-implementation through the OpenAI Gym (Brockman et al., 2016).

Table 4 checks our reinforcement learning baseline against the returns of the original DQN. Returns are better for all environments evaluated, justifying the baseline as reasonable for measuring further improvements due to self-supervision.

The code for the self-supervised tasks, policy optimization, and environment will be released.

	Pong	Qbert	Seaquest	S. Invaders	BeamRider	Breakout
NIPS DQN	20	1952	1705	581	4092	168
Our A3C	21	18028	1756	1102	5061	367

Table 4: Comparison of the best scores achieved by the original DQN (Mnih et al., 2013) and the same base architecture optimized with our A3C implementation. Training is carried out for 60M updates. Scores are reported as the mean of 100 runs with random no-op starts as in existing work. This sanity check demonstrates reasonable returns, so improvement from self-supervision cannot be attributed to deficiencies in the RL setup.

## 5 Related Work

Representation learning for reinforcement learning, robotics, and control is commonly known as state representation learning, as it yields the state for modeling the task as an MDP. That is, the goal of the state representation is to transform the history of observations, actions, and rewards into a sufficient state for efficient policy learning. This can be summarized formally as seeking a mapping  $\phi$  such that the current state  $s_t = \phi(o_{1:t}, a_{1:t}, r_{1:t})$  as in Jonschkowski & Brock (2015).

Unsupervised learning by auto-encoding is a common approach to state representation learning. The embed to control objective (Watter et al., 2015) combines variational auto-encoding with one-step dynamics modeling for image observations and locally-linear latent dynamics. The deep spatial auto-encoder (Finn et al., 2016) maps image observations into low-dimensional spatial coordinates by auto-encoding with a smoothness prior on the latent representation. The joint inverse and forward model of Agrawal et al. (2016) learns to poke objects by self-supervising inverse dynamics while predicting future states (not observations) for regularization. These approaches optimize policies to achieve a goal state without a task reward, so it is not possible to fine-tune the representation to optimize return. In contrast our auxiliary, discriminative losses capture dynamics, inverse dynamics, and other aspects of the environment in tandem with RL.

For deep RL, the use of pre-training and transfer is limited. ML-DDPG (Munk et al., 2016) extends actor-critic with a one-step predictive model of the successor state and reward. The observation mapping is learned by the first layer of the the model, transferred to the actor-critic network, and then fixed. Our successor self-supervision is discriminative rather than generative and we transfer all layers to the actor-critic network for end-to-end optimization. The end-to-end visuomotor policies of Levine et al. (2016) have the first layer initialized from supervised pre-training on ImageNet. Instrumented pose estimation pre-training further scaffolds the representation for policy optimization. Our auxiliary losses are purely self-supervised and only require regular transitions.

The robotic priors of Jonschkowski & Brock (2015) are auxiliary losses for temporal coherence, repeatability, proportionality, and causality. Multi-task optimization of these losses defines a linear, low-dimensional observation mapping for RL. These losses are defined on distances between states conditioned on action and reward, while we define discriminative losses on the  $(s, a, r, s')$  of transitions.

Concurrent work explores different methods to augment reinforcement learning with auxiliary losses. Jaderberg et al. (2016) extend value function estimation with instantaneous reward prediction using replay memory and introduce off-policy pseudo-reward control tasks. Mirowski et al. (2016) extend navigation tasks with auxiliary losses for spatial and path representations through coarse depth regression and a kind of loop closure for recognizing paths that have been already visited. Dosovitskiy & Koltun (2016) learn to predict future measurements of supervised, task-specific quantities such as the presence of enemies and health in a combat game (DOOM). In the same spirit as our work, these approaches seek to improve policy returns, data efficiency, and robustness of end-to-end RL. Our self-supervised tasks do not require additional privileged information, we focus on discriminative formulations of auxiliary losses, and we compare a variety of ambient signals for self-supervision.

## 6 Discussion

It is encouraging that self-supervision, with and without reward, can improve reinforcement learning. Pre-training alone suffices to show improvements especially during early iterations. Joint training further improves data efficiency by tracking the policy distribution during optimization.

Representation learning by self-supervision alone is agnostic to any particular task, and acts as a policy scaffold no matter the reward. This scaffold can be developed in the absence of an extrinsic reward whenever the policy is at play in the environment. A next step is to cast these losses into intrinsic rewards to further guide optimization. By augmenting RL with self-supervision, transitions without reward need not be so unrewarding for the representation.

## Acknowledgements

This work was supported in part by Berkeley AI Research, Berkeley Deep Drive, NSF, DARPA, NVIDIA, and Intel. We gratefully acknowledge NVIDIA for GPU donation. We thank John Schulman and Chelsea Finn for advice and useful discussions. We thank Alec Radford for sharing his implementation of A3C used in our joint optimization experiments. We thank Jeff Donahue for the care and feeding of the BiGANs.

## References

- Agrawal, Pulkit, Nair, Ashvin, Abbeel, Pieter, Malik, Jitendra, and Levine, Sergey. Learning to poke by poking: Experiential learning of intuitive physics. In *NIPS*, 2016.
- Barto, Andrew G, Singh, Satinder, and Chentanez, Nuttapon. Intrinsically motivated learning of hierarchical collections of skills. In *CDL*, pp. 112–119, 2004.
- Brockman, Greg, Cheung, Vicki, Pettersson, Ludwig, Schneider, Jonas, Schulman, John, Tang, Jie, and Zaremba, Wojciech. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.
- Chentanez, Nuttapon, Barto, Andrew G, and Singh, Satinder P. Intrinsically motivated reinforcement learning. In *NIPS*, pp. 1281–1288, 2004.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Doersch, Carl, Gupta, Abhinav, and Efros, Alexei A. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.
- Donahue, Jeff, Krähenbühl, Philipp, and Darrell, Trevor. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- Dosovitskiy, Alexey and Koltun, Vladlen. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.
- Finn, Chelsea, Tan, Xin Yu, Duan, Yan, Darrell, Trevor, Levine, Sergey, and Abbeel, Pieter. Deep spatial autoencoders for visuomotor learning. In *ICRA*, 2016.
- Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *NIPS*, pp. 2672–2680. 2014.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *CVPR*, 2016.

- Hinton, G.E. and Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507, 2006.
- Houthoofd, Rein, Chen, Xi, Duan, Yan, Schulman, John, De Turck, Filip, and Abbeel, Pieter. Variational information maximizing exploration. In *NIPS*, 2016.
- Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David, and Kavukcuoglu, Koray. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.
- Jonschkowski, Rico and Brock, Oliver. Learning state representations with robotic priors. *Autonomous Robots*, 39(3): 407–428, 2015.
- Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. In *ICLR*, 2014.
- Krähenbühl, Philipp, Doersch, Carl, Donahue, Jeff, and Darrell, Trevor. Data-dependent initializations of convolutional neural networks. In *ICLR*, 2016.
- LeCun, Yann A, Bottou, Léon, Orr, Genevieve B, and Müller, Klaus-Robert. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 1998.
- Levine, Sergey, Finn, Chelsea, Darrell, Trevor, and Abbeel, Pieter. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.
- Mirowski, Piotr, Pascanu, Razvan, Viola, Fabio, Soyer, Hubert, Ballard, Andy, Banino, Andrea, Denil, Misha, Goroshin, Ross, Sifre, Laurent, Kavukcuoglu, Koray, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. Playing atari with deep reinforcement learning. *arXiv*, 2013.
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- Mnih, Volodymyr, Badia, Adrià Puigdomènech, Mirza, Mehdi, Graves, Alex, Lillicrap, Timothy P, Harley, Tim, Silver, David, and Kavukcuoglu, Koray. Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*, 2016.
- Munk, Jelle, Kober, Jens, and Babuška, Robert. Learning state representation for deep actor-critic control. 2016.
- Noroozi, Mehdi and Favaro, Paolo. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.
- Oudeyer, Pierre-Yves, Kaplan, Frédéric, Hafner, Verena V, and Whyte, Andrew. The playground experiment: Task-independent development of a curious robot. In *AAAI Spring Symposium on Developmental Robotics*, pp. 42–47. Stanford, California, 2005.
- Pathak, Deepak, Krähenbühl, Philipp, Donahue, Jeff, Darrell, Trevor, and Efros, Alexei. Context encoders: Feature learning by inpainting. 2016.
- Rezende, Danilo J, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pp. 1278–1286, 2014.
- Schmidhuber, Jürgen. Curious model-building control systems. In *IJCNN*, pp. 1458–1463. IEEE, 1991.
- Sutton, Richard S and Barto, Andrew G. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Watter, Manuel, Springenberg, Jost, Boedecker, Joshka, and Riedmiller, Martin. Embed to control: A locally linear latent dynamics model for control from raw images. In *NIPS*, pp. 2746–2754, 2015.

Zhang, Richard, Isola, Phillip, and Efros, Alexei A. Colorful image colorization. In *ECCV*, 2016.