Learning to Play with Intrinsically-Motivated Self-Aware Agents

Nick Haber [1] **Damian Mrowca** [*2] **Li Fei-Fei** [2] **Daniel L. K. Yamins** [1 2]

Abstract

Infants are experts at playing, with an amazing ability to generate novel structured behaviors in unstructured environments that lack clear extrinsic reward signals. We seek to mathematically formalize these abilities using a neural network that implements curiosity-driven intrinsic motivation. Using a simple but ecologically naturalistic simulated environment in which an agent can move and interact with objects it sees, we propose a "world-model" network that learns to predict the dynamic consequences of the agent's actions. Simultaneously, we train a separate explicit "self-model" that allows the agent to track the error map of its own world-model, and then uses the self-model to adversarially challenge the developing world-model. We demonstrate that this policy causes the agent to explore novel and informative interactions with its environment, leading to the generation of a spectrum of complex behaviors, including ego-motion prediction, object attention, and object gathering. Moreover, the world-model that the agent learns supports improved performance on object dynamics prediction, detection, localization and recognition tasks. Taken together, our results are initial steps toward creating flexible autonomous agents that self-supervise in complex novel physical environments.

1. Introduction

Truly autonomous artificial agents must be able to discover useful behaviors in complex environments without having humans present to constantly pre-specify tasks and rewards. This ability is beyond that of today's most advanced autonomous robots. For example, NASA's Curiosity Rover can only explore the Mars with a few pre-configured task programs. This severely limits Curiosity's long-term utility, as it cannot set itself new tasks that will help it learn to take better advantage of the Martian environment over time.

[*]Equal contribution [1]Department of Psychology / [2]Computer Science, Stanford University, Stanford, CA 94305, USA. Correspondence to: Nick Haber <nhaber@stanford.edu>, Damian Mrowca <mrowca@stanford.edu>

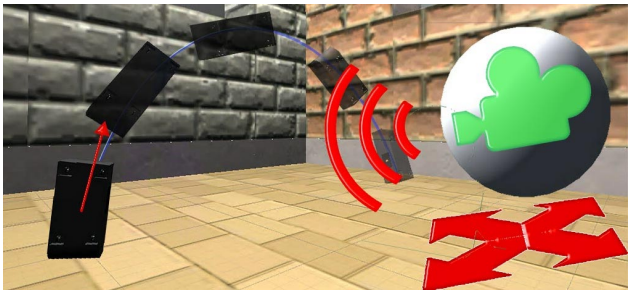


Figure 1. An agent is embedded in a three-dimensional environment where it can move around, apply forces to visible objects in close proximity, and receive visual input. What policies allow the agent to learn a general-purpose world-model?

In contrast, human infants exhibit a wide range of interesting, apparently spontaneous, visuo-motor behaviors — including navigating their environment, seeking out and attending to novel objects, and engaging physically with these objects in novel and surprising ways (Fantz, 1964; Twomey & Westermann, 2017; Hurley et al., 2010; Hurley & Oakes, 2015; Goupil et al., 2016; Begus et al., 2014; Gopnik et al., 2009). In short, young children are excellent at playing — "scientists in the crib" (Gopnik et al., 2009) who create, intentionally, events that are new, informative, and exciting to them (Sokolov, 1963; Fantz, 1964). Aside from being fun, play behaviors are an active learning process (Settles, 2011), driving the self-supervised learning of representations underlying sensory judgments and motor planning capacities (Kidd et al., 2012; Goupil et al., 2016; Begus et al., 2014).

But how can we use these observations on infant play to improve artificial intelligence? AI theorists have long realized that playful behavior in the absence of rewards can be mathematically formalized via loss functions encoding intrinsic reward signals, in which an agent chooses actions that result in novel but predictable states that maximize its learning (Schmidhuber, 2010). These ideas rely on a virtuous cycle in which the agent actively self-curricularizes as it pushes the boundaries of what its world-model-prediction systems can achieve. As world-modeling capacity improves, what used to be novel becomes old hat, and the cycle starts again.

Here, we build on these ideas using the tools of modern deep reinforcement learning to create an artificial agent that learns to play. We construct a simulated interactive physical environment in which an agent can move around and

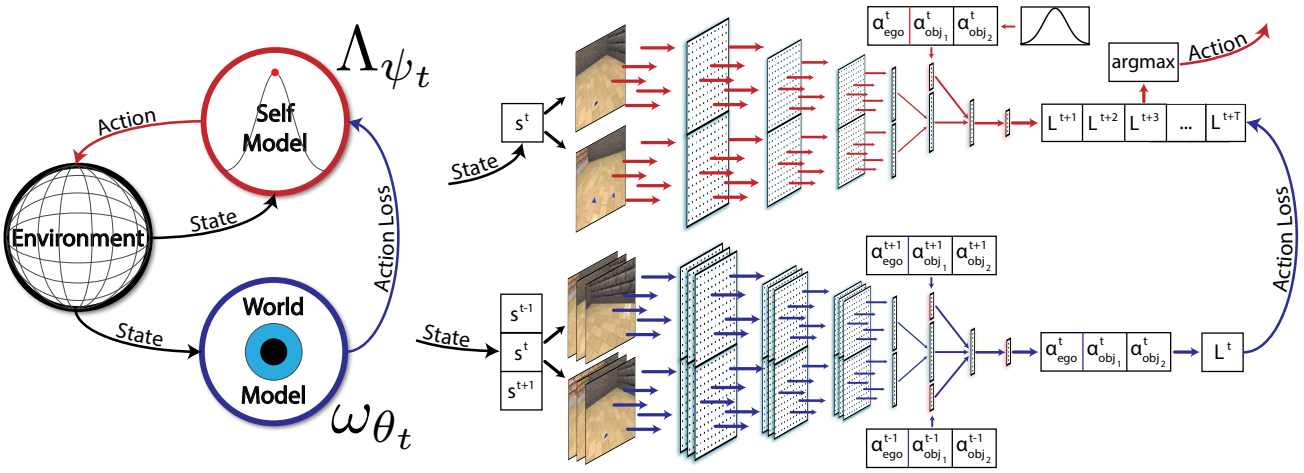


Figure 2. Intrinsically-motivated self-aware agent architecture. The world-model (blue) solves a dynamics prediction problem. Simultaneously a self-model (red) is learned that seeks to predict the world-model's loss. Actions are chosen to antagonize the world-model, leading to novel and surprising events in the environment (black).

physically act on objects it sees (Fig. 1). In this world, interesting interactions are possible, but sparse unless actively sought. We then describe a neural network architecture through which the agent learns a world-model that predicts the consequences of agent's actions, either through forward or inverse dynamics prediction. As the agent optimizes the accuracy of its world-model, a separate explicit "self-model" neural network simultaneously learns to predict the errors of the agent's own world-model. Based on the self-model, the agent then uses an action policy that seeks to take actions that adversarially challenge the current state of its world-model. We demonstrate that this intrinsically-motived self-aware architecture stably engages in the virtuous reinforcement learning cycle described above, spontaneously learning to understand self-generated ego-motion and to selectively pay attention to, localize, recognize, and interact with objects, without having any of these concepts built in. This learning occurs through an emergent active self-supervised process in which new capacities arise at distinct "developmental milestones" like those in human infants. These results are steps toward creating mathematically well-motived, flexible autonomous agents that use intrinsic motivation to learn about and spontaneously generate useful behaviors for adapting to unknown environments.

2. Related Work

Our work connects to a variety of existing ideas in self-supervision, active learning, and deep reinforcement learning. At the most basic level, auto-encoders can develop representations by reconstructing input images (Olshausen & Field, 1997; Kingma & Welling, 2013). More explicit self-supervised auxiliary tasks include semantic segmentation (Hong et al., 2017), pose estimation (Mitash et al., 2017), solving jigsaw puzzles (Noroozi & Favaro, 2016), colorization (Zhang et al., 2016), and rotation (Spyros Gidaris, 2018). Self-supervision on videos in the form of future frame prediction may have potential to surpass the performance of the aforementioned methods (Kalchbrenner et al., 2017), but a challenge facing frame prediction is that most sequences in recorded videos are "boring", with little interesting dynamics occurring from one frame to the next.

In order to encourage interesting events to happen, it is useful for the agent to have the capacity to interact with its environment, or at least to select the data that it sees in training. In traditional active learning, an agent seeks to learn a supervised task from using little labeled data as possible, with the ability to request more labeled data if necessary (Settles, 2011; Gilad-Bachrach et al., 2005). Recent optimization methods trade-off uncertainty and diversity to obtain diversified sets of hard examples (Elhamifar et al., 2013; Sener & Savarese, 2017), or use heuristics to assign labels to data examples with high confidence while querying labels for examples with low confidence (Wang et al., 2017).

Going beyond selection of examples from a pre-determined set, recent work in robotics (Agrawal et al., 2016; Popov et al., 2017) study learning tasks in interactive visuo-motor environments. In particular, Finn & Levine (2017) and Ebert et al. (2017) have tried to learn self-supervised visuo-motor tasks with robot arms. The results are promising but suffer from the challenges of having to predict forward dynamics in pixel space and having to orchestrate random pushing motions to generate training data. These works do not use an intrinsically driven mechanism that would bias the robot to explore its environment in a structured way.

Intrinsic and extrinsic reward structures have been used to learn generic options for a variety of tasks (Chentanez et al., 2005; Singh et al., 2010). Houthooft et al. (2016)

demonstrated that reasonable exploration-exploitation trade-offs can be achieved by intrinsic reward terms formulated as information gain. Frank et al. (2014) use information gain maximization to implement artificial curiosity on a humanoid robot. Kulkarni et al. (2016) combine intrinsic motivation with hierarchical action-value functions operating at different temporal scales, for goal-driven deep reinforcement learning. Achiam & Sastry (2017) formulate surprise for intrinsic motivation as the KL-divergence of the true transition probabilities from learned model probabilities. Held et al. (2017) use a generator network, which is optimized using adversarial training to produce tasks that are always at the appropriate level of difficulty for an agent, to automatically produce a curriculum of navigation tasks to learn. Jaderberg et al. (2016) show that target tasks can be improved by using auxiliary intrinsic rewards.

Closest to our work in its formulation of an intrinsic reward signal is Pathak et al. (2017). Their work uses curiosity to antagonize a future prediction signal in the latent space of a inverse dynamics prediction task to improve learning in video games, showing that intrinsic motivation leads to faster floor-plan exploration in a 2D game environment. Our work differs from theirs in using a physically realistic three-dimensional environment, and shows how in this context, intrinsic motivation can lead to substantially more sophisticated agent-object behavior generation (the "playing"). We also show the learned representation transfers to improved performance on analogs of real-world visual tasks, such as object localization and recognition. Underlying the difference of our technical approach is our introduction of an explicit self-model, representing the agent's awareness of its own internal state. This difference can be viewed in RL terms as the use of a more explicit model-based architecture (ours) in place of a model-free setup. To our knowledge, a self-supervised setup in which an explicitly self-modeling agent uses intrinsic motivation to learn about and restructure its environment has not been explored prior to this work.

3. Interactive Physical Environment

Our agent is situated in a physically realistic simulated *environment* (black in Fig. 2) built in Unity 3D (Fig. 1) along with several objects. These objects interact according to Newtonian physics as simulated by the PhysX engine. The agent's avatar is a sphere that swivels in place, moves around, and receives RGB images from a forward-facing camera (as in Fig. 1). The agent can apply forces and torques in all three dimensions to any object(s) that are both in view and within a fixed distance δ of the agent's position. Although the floor and walls of the environment are static, the agent and objects can collide with them. The action space of the agent is a subset of $\mathbb{R}^{2+6N}$, in which the first 2 dimensions specify ego-motion, restricting agent movement to forward/backward motion v_{fwd} and horizontal planar rotation v_θ, while the remaining $6N$ dimensions specify the forces f_x, f_y, f_z and torques $\tau_x, \tau_y \tau_z$ applied to N objects sorted from the lower-leftmost to the upper-rightmost object relative to the agent's field of view. All coordinates are bounded by constants and normalized to 1.

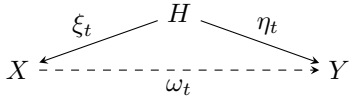
4. Agent Architecture

Our agent consists of a *world-model* and a *self-model* (Fig. 2). The world-model is tasked to learn a dynamics prediction problem based on inputs from the environment. The self-model tries to estimate the world-model's losses several time steps into the future, as a function of potential agent actions. An action choice policy based on the self-model then chooses actions that antagonize the world-model's learning. In this section, we formalize these ideas mathematically.

4.1. World-Model

Figuring out a tractable dynamics prediction problem to make the target of intrinsic motivation is an important first challenge. We begin with an abstract mathematical treatment to expose the key issues. Let (S, P, O, A) define a partially observable Markov Decision Process (POMDP) with state space S, transition dynamics P, observations O, and action space A but no specified external reward. In the physics environment described above, states S encode the positions, velocities, and physical properties of an agent and objects in a 3-D physical space. Dynamics P are the (deterministic) updates given by Newtonian physics. Observations O are the images rendered by an agent-mounted camera. Actions A encode self-motions of the agent as well as and forces/torques that the agent can apply to objects.

Within this context, agents make decisions about what action to take at each time, accumulating histories of state-action pairs $\{(s_0, a_0), \ldots, (s_t, a_t), \ldots\}$. Let H be the set of (fixed-length) windows in such histories. Informally, a *dynamics prediction problem* is a pairing of complementary subsets of data — "inputs" and "true values" — generated from H. The goal of the agent is to learn the map from inputs to true-values. Mathematically, we define this as a five-tuple (X, Y, ξ_t, η_t, L), where X is the input space, Y is the true-value space, ξ_t is a (possibly time-varying) map from histories to inputs, and η is a (possibly time-varying) map from histories to true-values, and L is a "loss function" $L : Y \times Y \to \mathbb{R}^{\geq 0}$ such that $L(y, y) = 0$ for $y \in Y$:



The agent is equipped with a time-evolving *world-model*, which is a map $\omega_{\theta_t} : X \to Y$, defined by learnable parameters θ_t. The agent attempts to evolve θ_t over time so that $L_\omega = L(\omega_{\theta_t}(\xi_t(h)), \eta_t(h))$ is minimized. In words,

the agent's world-model (blue in Fig. 2) tries to learn to reconstruct the missing true-value from the input datum. This is done regardless of whether ξ_t and η_t actually induce a well-defined mapping $X \to Y$ that makes the diagram above commute. Obstructions to the existence of such a commuting map arise from the *degeneracy* of the system dynamics P (situations where the same input corresponds to two different true-values).

The most natural dynamics problem is *forward dynamics prediction*. For notational convenience, for any historical observable x, let $x^t_{k_0:k_1}$ for $k_0 \leq k_1$ denote the sequence of values $(x_{t+k_0}, \ldots, x_{t+k_1})$. For b steps in the past and f steps in the future, let $h^t = (s^t_{-b:f}, a^t_{-b:f-1})$ define the temporal observation window available to the agent. Forward dynamics prediction is now defined by letting $\xi_F(h^t) = (o^t_{-b:0}, a^t_{-b:f-1})$ and $\eta_F(h^t) = o^t_{1:f}$, where o_t is the observation corresponding to state s_t. In other words, the agent is trying to predict the next (several) observation given past observations, past actions, and its current action. In real 3-D physical domains, the true-values $o^t_{1:f}$ correspond to f bitmap image arrays of future frames, and the loss function L_F is either ℓ_2 loss on pixels, or some discretization thereof. Despite recent progress on the frame prediction problem (Kalchbrenner et al., 2017; Finn & Levine, 2017), it remains quite challenging, in part because the dimensionality of the true-value space is so large.

In practice, it can be substantially easier to solve the *inverse dynamics prediction*. This is defined by $\xi_I(h^t) = (o^t_{-b:f}, a^t_{-b:-1}, a^t_{1:f-1})$ and $\eta_I(h^t) = a_t$. In other words, the agent is trying to "post-dict" which current action was needed to have generated the observed sequence of observations, given knowledge of its past and future actions. Here, the loss function L_I is computed on low-dimensional action space, a problem that has proven tractable (Agrawal et al., 2016). However, it can suffer from substantial degeneracy: consider the case of an agent pressing an object downward into the ground. No matter what the force downward applied is, the object does not move, so input information in X (the sequence of object positions) is insufficient to determine what the true-value in Y (the action) was.

A more sophisticated concept that tries to solve both high-dimensionality and degeneracy simultaneously is *latent space future prediction* (Pathak et al., 2017). In this case, we begin with a system solving the inverse dynamics prediction problem, and assume that its parametrization of world-models factor into a composition $\omega^I_{\theta_t} = d_{\beta_t} \circ e_{\alpha_t}$ where α_t and β_t are non-overlapping sets of parameters. In this case, we call $e_t = e_{\alpha_t}$ the *encoding*, $d_t = d_{\alpha_t}$ the *decoding*, and the range of e_t the *latent space* $\mathcal{L}$ of the problem. Now, we define (time-varying) ξ^t_{LF}, η^t_{LF} as the 1-time-step future prediction problem on trajectories in $\mathcal{L}$ given by the time-varying encoding, i.e. by $\xi^t_{LF}(h^t) = (e_t(o^t)_{-b:0}, a^t_{-b:0})$ and $\eta^t_{LF}(h^t) = e_t(o_{t+1})$. The inverse-prediction world-model ω_I and latent-space world-model ω_{LF} evolve separately but simultaneously. If $\mathcal{L}$ is sufficiently low dimensional, this may be a good compromise task.

In this work, we explore both inverse dynamics and latent space future prediction as world-model tasks.

4.2. Explicit Self-Model

The agent's goal is to antagonize its world-model, so if it could predict L_{ω_t} incurred at future time steps as a function of its current action, a simple antagonistic policy could seek to maximize L_{ω^t} over some number of time steps in the future. Given s_t and a proposed next action a, the self-model Λ (red in Fig. 2) predicts

$$\Lambda(s_t, a_t) = \mathbf{p}(c \mid s_t, a) \in \prod_{t=1}^{T} \mathcal{P}([C]), \tag{1}$$

probability distributions over C discrete (via thresholding) classes of world-model loss for a set number T of future time steps. It is penalized with a softmax cross-entropy loss. Note that all future losses aside from the first one, depend not only on the state of the world-model, but also on future actions taken, and the self-model hence needs to predict in expectation over future policy.

In the context of a 3-D physical environment, loss predictions can be interpreted as *self prediction maps* $\Lambda_{s_t}[a]$ over action space given a current state s_t. This interpretation is useful for intuitively visualizing what strategy the agent is taking in any given situation (see Fig 4).

4.3. Adversarial Action Choice Policy

Given the self-model, the agent can use a simple mechanism to choose its actions. The self-model provides, given s_t and a proposed next action a, a map

$$S \times A \to \prod_{t=1}^{T} \mathcal{P}([C]). \tag{2}$$

Given an additional summary mapping

$$\sigma : \mathrm{map}(A, \prod_{t=1}^{T} \mathcal{P}([C])) \to \mathrm{map}(A, \mathbb{R}), \tag{3}$$

this provides us a real-valued map $a \mapsto \sigma_\Lambda(a) = \sigma(\Lambda(s_t, a))$, which then allows us to define a probability distribution on the next action chosen

$$\pi(a) \sim \exp(\beta \sigma_\Lambda(a)) \tag{4}$$

with given hyperparameter β. In what follows, we use as σ sum over expectation values for each time step, although more sophisticated functions to combine rewards over time are possible (Schulman et al., 2017). In practice, we execute our policy by evaluating $L_{s_t, a_{t-1}}(a)$ for K uniform random samples in A. We then sample from a K-way discrete distribution with probabilities proportional to eq. (4).

4.4. Models and Metrics of Comparison

We use convolutional neural networks as the base architecture to learn both world-models ω_θ and self-models Λ_ψ. In the specific experiments described below, these networks have an encoding structure with a common architecture involving twelve convolutional layers, two-stride max pools every other layer, and one fully-connected layer, to encode all states into a lower-dimensional latent space L, with shared weights across time. For the inverse dynamics task, the top encoding layer of the network is combined with actions $\{a_{t'} \mid t' \neq t\}$, fed into a two-layer fully-connected network, on top of which a softmax classifier is used to predict action a_t. For the latent space future prediction task, the top encoding layer is used as the latent space $\mathcal{L}$, and the latent model ω_{LF} is estimated with another copy of the encoding network. For the self-model, the top encoding layer is combined with action a_t, and then fed into a two-layer fully connected network to predict world-model loss with that action. Parameters θ, ψ are trained end-to-end by stochastic gradient descent from randomly initialization.

We compare a variety of agents defined by different combinations of world-model task and policy mechanism. Several baseline models include the inverse dynamics (ID) problem with a random action policy (ID-RP), the ID problem with random-weight encoding and random policy (IDRW-RP), and the ID problem with random-weight encoding and self-model based policy (IDRW-SP). These baselines are compared to more powerful agents with a fully learnable encoding and the self-model policy, both for the ID problem (ID-SP), and the latent space future prediction (LF-SP).

For each of these on comparison models, we evaluate three types of metrics: **(1)** **Dynamics prediction tasks** where we measure the inverse dynamics prediction performance on two held-out validation subsets of data: (i) an *easy* dataset drawing from the uncontrolled background distribution of events (i.e. the random policy), dominated by ego-motion; and (ii) a *hard* dataset that is enriched for events that we have observed to be challenging — e.g. frames in which one or more object is present. This metric measures active learning gains, assessing to what extent the agent self-constructs training data for the hard subset while retaining performance on the easy dataset. We also look at **(2)** **emergent behavior**, quantifying the appearance of interesting behaviors such as attention to and acting on objects (as opposed to mere self-motion), navigation and planning, and ability to cause multiple objects to interact. We track not only how much time the agent spends playing with an object, but also the relationship between when this behavior appears and other observables, such as sharp changes in overall world-model loss. Finally, we measure **(3)** **task transfer.** including the ability of the agent model to predict object presence, location, and category identity.

5. Experiments

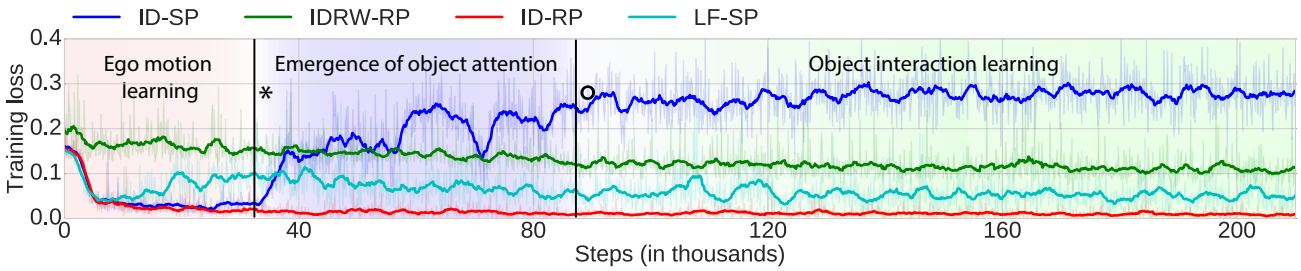
We randomly place the agent and up to two objects in a square 10 by 10 units room. In total, we train the agent on 16 blue objects with different shapes, i.e. cones, cylinders, cuboids, pyramids, and spheroids of varied aspect ratios. We gather data using 16 simulation instances asynchronously with different seeds and objects. We reinitialize the scene every 2^{13} to 2^{15} steps. Each simulation maintains a buffer of 250 time steps. For model updates 2 examples are randomly sampled from each simulation buffer to form a batch of size 32. We train our models using the Adam algorithm (Kingma & Ba, 2014), with a learning rate of 0.0001.

We first place the agent into a room with one object, and evaluate its ability to predict inverse dynamics, and attend to, localize, recognize and navigate towards objects.

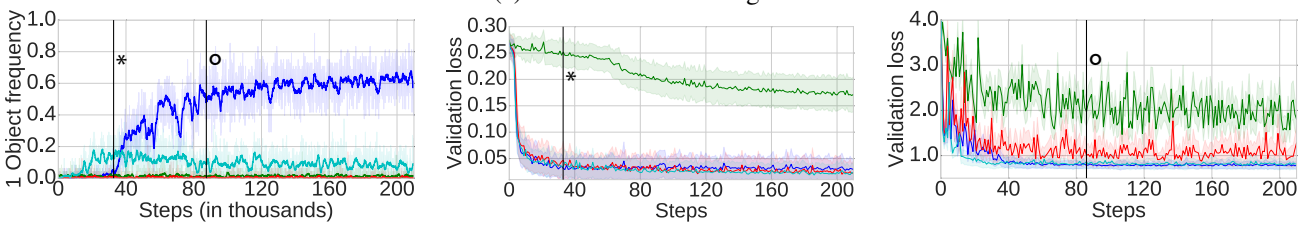
Ego-motion learning. Fig. 3 (a) shows the total training loss curves of the ID-SP, LF-SP models and baselines. The random-encoding IDRW-RP model (green) learns ineffectively on the background random data distribution. All other models learn ego-motion prediction effectively. The ID-RP model quickly converges to a low loss value, where it remains from then on, having effectively learned ego-motion prediction without an antagonistic policy since ego-motion interactions are common in the background random data distribution. The ID-SP and LF-SP models also learn ego-motion effectively, as seen in the initial decrease of their loss seen in Fig. 3 (a), and in the low final loss in Fig. 3 (c) which depicts the easy ego-motion validation dataset. The ego-motion accuracy reported in Table 1 is close to the total validation accuracy reached at this point.

Emergence of object attention. For both learned-weight agents implementing SP, loss increases after an initial decrease due to ego-motion learning, as seen in Fig. 3 (a). As shown in Fig. 3 (b), this loss increase corresponds to emergence of object attention. Both ID-SP and LF-SP agent exhibit increased frequency of frames with an object present, coinciding with the increase in world-model loss, though in the ID-SP agent this is substantially more pronounced. After convergence, the ID-SP agent is interacting with objects about $60\,\%$ of the time. Baselines models almost never interact with the object. World-model loss increases for the self-model driven agents, since object interactions are much harder to predict than simple ego-motions.

Navigation and planning. SP agents also exhibit navigation and planning abilities. In Fig. 4 we give visualizations of self prediction maps projected onto the agent's position at the respective time. The maps are generated by uniformly sampling 1000 actions a from the action space A, evaluating $\Lambda_{s_t}(a)$ and applying a post-processing smoothing algorithm. We show an exemplary sequence of 12 time steps. The self prediction maps show the agent predicting a higher loss



(a) World-model training loss



(b) Object frequency (c) Easy dataset (ego-motion) (d) Hard dataset (object interactions)

Figure 3. 1 object experiments. (a) World-model training loss. (b) Percentage of frames in which an object is present. (c) World-model test-set loss on "easy" ego-motion-only data, with no objects present. (d) World-model test-set loss on "hard" validation data, with object present, where agent must solve object physics prediction.

(red) for actions moving it towards the object to reach a play state. As a result, the self-model driven agents take actions to navigate closer to the object.

Improved object dynamics prediction. Object attention and navigation lead SP agents to substantially different data distributions than baselines. We evaluate the inverse dynamics prediction performance on the held out hard object interaction validation set. Here, the ID-SP and LF-SP agents outperform the baselines on predicting the harder object interaction subset by a significant margin, showing that increased object attention translates to improved inverse dynamics prediction (see Fig. 3 (d) and Table 1). Crucially, even though ID-SP and LF-SP have substantially decreased the fraction of time spent on ego-motion interactions as was previously observed in Fig. 3 (c), they still retain high performance on that "easier" task.

Improved task transfers. We next measure the agent's abilities to solve tasks for which they were not directly trained, including object presence, localization, and recognition. We build linear estimators on learned features from each agent world-model, trained on off-line validation datasets consisting of 16,000 image pairs labeled, respectively, with the object's presence, its pixel-wise 2D centroid position, or 16-way object category. Results are reported on test sets comprising 8,000 image pairs each. Note that the test sets contain substantial variation in position, pose and size, rendering these tasks nontrivial. Self-model driven agents substantially outperform alternatives on all three transfer tasks, As shown in Table 1, the SP ($T = 5$) agent outperforms baselines on inverse dynamics and object presence metrics, while ID-SP outperforms LF-SP on localization and recognition.

Emergence of multi-object interactions. In a second experiment we increase the number of objects to two. At the beginning of the training of the two object experiment we observe similar stages as for the one object experiment (Fig. 5 (a)) for both ID-SP and LF-SP. The loss temporarily decreases as the agent learns to predict its ego-motion and rises when its attention shifts towards objects which it then interacts with. For ID-SP agents with sufficiently long time horizon (e.g. $T = 40$), we robustly observe the emergence of an additional stage in which the loss increases further corresponding to the agent gathering and playing with two objects simultaneously. This is reflected in an increase in two object play time (Fig. 5 (c)) over one object play time (Fig. 5 (b)). Moreover, the average distance between the agent and the objects decreases over time as seen in Fig. 5 (c). We do not observe this additional stage either for ID-SP shorter time horizon (e.g. $T = 5$) or for the LF-SP model even with longer horizons (e.g. $T = 40$). Unsurprisingly, the ID-RP baseline with its random policy experiences a quick loss drop and flattening out of loss. The ID-SP agent has discovered how to take advantage of the increased difficulty and therefore "interestingness" of two object configurations (compare blue with green horizontal line in Fig. 5 (a)). Interestingly, we find that training with two objects present improves recognition transfer performance as compared to one object scenarios, potentially due to the greater complexity of two-object configurations (Table 1). This is especially notable for the ID-SP ($T = 40$) agent which constructs a substantially increased percentage of two-object events.

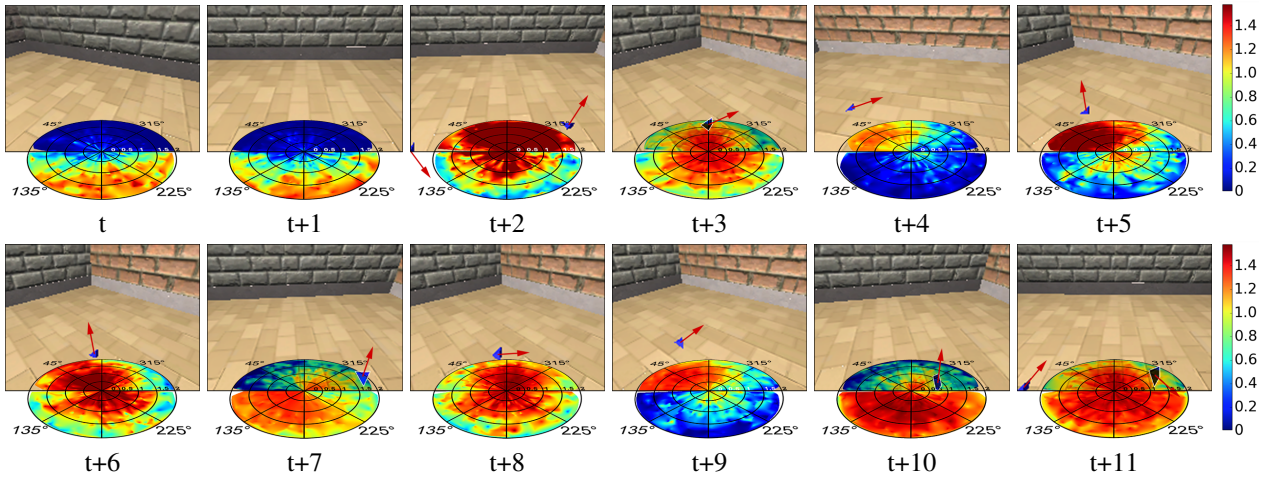


Figure 4. Navigation and planning behavior. Exemplary model roll-out for 12 consecutive time steps. Red force vectors on the objects depict the predicted actions that would maximize the world-models loss. Ego-motion self prediction maps are drawn at the center of the agents position. Red colors correspond to high and blue colors to low loss predictions. The agent starts off without seeing an object and predicts higher loss if it turns around to explore for an object. Once an object is in view the self-model predicts higher loss if the agent approaches an object if it is far away, or turns towards an object to keep it in view once it is close.

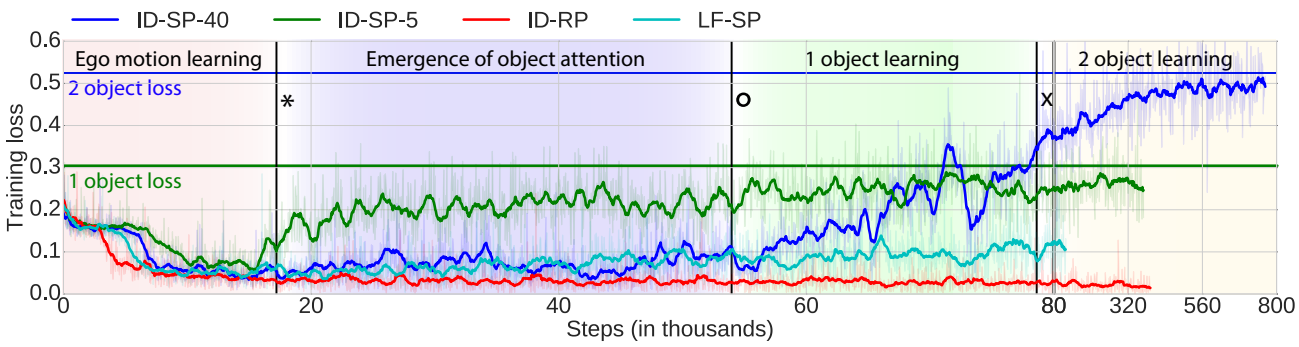
Table 1. Performance comparison. Ego-motion (v_{fwd}, v_θ) and interaction (f, τ) accuracy in % is compared for play and non-play states. Object frequency, presence and recognition are measured in % and localization in mean pixel error. All models were trained with one object per room unless otherwise stated.

TASK	IDRW-RP	IDRW-SP	ID-RP	ID-SP	LF-SP
v_{fwd} ACCURACY — EASY	65.9	56.0	**96.0**	95.3	95.3
v_θ ACCURACY — EASY	82.9	75.2	**98.7**	98.4	98.5
v_{fwd} ACCURACY — HARD	62.4	69.2	90.4	**95.9**	95.4
v_θ ACCURACY — HARD	79.0	80.0	95.5	**98.2**	**98.1**
f ACCURACY — HARD	20.8	33.1	42.1	**51.1**	45.1
τ ACCURACY — HARD	20.9	32.1	41.3	**43.2**	**43.2**
OBJECT FREQUENCY	0.50	47.9	0.40	**61.1**	12.8
OBJECT PRESENCE ERROR	4.0	3.0	0.92	0.92	**0.60**
LOCALIZATION ERROR [PX]	15.04	10.14	5.94	**4.36**	5.94
RECOGNITION ACCURACY	13.0	21.99	12.3	**28.5**	18.7
RECOGNITION ACCURACY – 2 OBJECT TRAINING	12.0	-	16.1	**39.7**	21.1

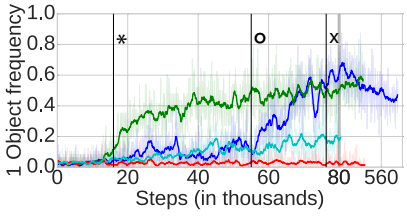
6. Discussion

We have constructed a simple and general intrinsic motivation mechanism in which a physically-embedded agent makes a world-model, then explicitly creates a "self-aware" meta-model of its own world-model, and then uses this self-model to adversarially antagonize the world-model. We have shown that this architecture allows an agent to spontaneously generate a spectrum of emergent naturalistic behaviors. Through self-curricularization in an active learning process the agent achieves several "developmental milestones" of suitably increasing complexity as it learns to "play". Starting with random actions, it quickly learns the dynamics of its own ego-motion. Then, without being given an explicit supervision signal as to the presence or location of an object, it discards ego-motion prediction as boring and begins to focus its attention on objects, which are more interesting. Lastly, when multiple objects are available, it gathers the objects so as to bring them into interaction range of each other. Throughout, the agent finds its way towards a more challenging data distribution that is at each moment just hard enough to expose the agent to new situations, but still understandable and exploitable by the agent. This intrinsically motived policy leads to performance gains in its understanding of object dynamics and other useful tasks for which the system did not receive an explicit training signal.

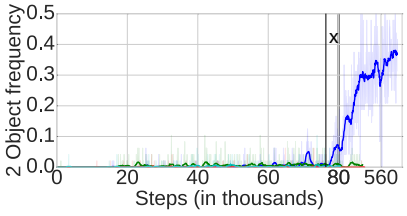
This occurs without any pre-trained visual backbone — the world-model was intentionally not initialized with filter weights pre-trained on (e.g.) ImageNet classification. This result constitutes partial progress in replacing the training of a visual backbone through a task such as large-scale image classification with an interactive self-supervised task and is a proof-of-concept that more complex milestones can be po-



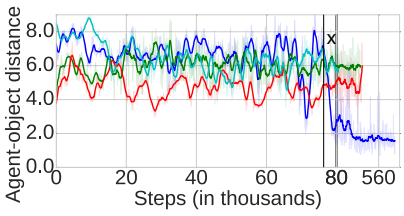
(a) World-model training loss



(b) One Object frequency



(c) Two Object frequency



(d) Average agent-object distance

Figure 5. 2 object experiments. (a) World-model training loss. (b) Percentage of frames in which one object is present. (c) Percentage of frames in which two objects are present. (d) Average distance between agent and objects in Unity units. For this average to be low (e.g. values of approximately two) both objects must be close to the agent simultaneously.

tentially reached while developing an understanding of object categories and physical relations. This combination of spontaneous behavior leading to an improved world-model is well suited to designing agents that must self-supervise in real-world reinforcement learning scenarios in which rewards are sparse or potentially unknown.

7. Future Work

A variety of limitations of the current work will need to be overcome in future work. First, to make our results better transfer to the real world, our environment and agent themselves need to be more realistic. On the one hand, better graphics and physics, with more varied and interesting visual objects, will be important to allow better transfer of learned behavior to real-world visuo-motor interactions. It will also be important to create a properly embodied agent with visible arms and tactile feedback, allowing for more realistic interactions. In addition, including other animate agents will not only lead to more complex interactions, but potentially also better learning through imitation (Ho & Ermon, 2016). In this scenario, the self-model component of our architecture will need to be not only aware of the agent itself, but also make predictions about the actions of other agents — connecting to what is known in cognitive science as *theory of mind* (Saxe & Kanwisher, 2003).

Second, the reinforcement learning techniques used should be improved to better handle more complex interactions beyond those demonstrated here. For interactions that are part of a larger experiment, e.g. placing an object on a table or a ramp and then watching it fall, sophisticated RL policies are likely to be necessary, with better ability to handle temporally extended reward schedules. It will also likely be necessary to use recurrent networks to meet working memory demands in such scenarios.

Third, our world-model needs to use better representations to improve at predicting such complex interactions. Our current approach suffers from *degenerate cases* in the inverse dynamics prediction problem — the problem does not correspond to a well-defined map. Though the latent space approach of (Pathak et al., 2017) is meant in part to ameliorate this issue, we have not yet found an entirely effective solution in our context. To resolve this issue, it will likely be key to both innovate on which dynamics prediction tasks to use for the world-model and better integrate their interaction with antagonistic action policies in the self-model.

Ultimately, by combining solutions to each of these challenges, we hope to build, and mathematically understand, substantially more effective autonomous agents.

Acknowledgements

This work was supported by the James S. McDonnell and Simons Foundations (DLKY), a Berry Foundation postdoctoral fellowship (NH), and the NVIDIA Corporation.

References

Achiam, Joshua and Sastry, Shankar. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017.

Agrawal, Pulkit, Nair, Ashvin V, Abbeel, Pieter, Malik, Jitendra, and Levine, Sergey. Learning to poke by poking: Experiential learning of intuitive physics. In *NIPS*, 2016.

Begus, Katarina, Gliga, Teodora, and Southgate, Victoria. Infants learn what they want to learn: Responding to infant pointing leads to superior learning. *PLOS ONE*, 9 (10):1–4, 10 2014. doi: 10.1371/journal.pone.0108817.

Chentanez, Nuttapong, Barto, Andrew G, and Singh, Satinder P. Intrinsically motivated reinforcement learning. In *NIPS*, pp. 1281–1288, 2005.

Ebert, Frederik, Finn, Chelsea, Lee, Alex X., and Levine, Sergey. Self-supervised visual planning with temporal skip connections. In *CoRL*, volume 78, pp. 344–356. PMLR, 2017.

Elhamifar, Ehsan, Sapiro, Guillermo, Yang, Allen Y., and Sastry, S. Shankar. A convex optimization framework for active learning. In *ICCV*, pp. 209–216. IEEE Computer Society, 2013. ISBN 978-1-4799-2839-2.

Fantz, R. L. Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146:668–670, 1964.

Finn, Chelsea and Levine, Sergey. Deep visual foresight for planning robot motion. In *ICRA*, pp. 2786–2793. IEEE, 2017. ISBN 978-1-5090-4633-1.

Frank, Mikhail, Leitner, Jrgen, Stollenga, Marijn F., Frster, Alexander, and Schmidhuber, Jrgen. Curiosity driven reinforcement learning for motion planning on humanoids. *Front. Neurorobot.*, 2014, 2014.

Gilad-Bachrach, Ran, Navot, Amir, and Tishby, Naftali. Query by committee made real. In *NIPS*, pp. 443–450, 2005.

Gopnik, A., Meltzoff, A.N., and Kuhl, P.K. *The Scientist In The Crib: Minds, Brains, And How Children Learn*. HarperCollins, 2009. ISBN 9780061846915.

Goupil, Louise, Romand-Monnier, Margaux, and Kouider, Sid. Infants ask for help when they know they dont know. *Proceedings of the National Academy of Sciences*, 113 (13):3492–3496, 2016. doi: 10.1073/pnas.1515129113.

Held, David, Geng, Xinyang, Florensa, Carlos, and Abbeel, Pieter. Automatic goal generation for reinforcement learning agents. *CoRR*, abs/1705.06366, 2017.

Ho, Jonathan and Ermon, Stefano. Generative adversarial imitation learning. In *NIPS*, pp. 4565–4573, 2016.

Hong, Seunghoon, Yeo, Donghun, Kwak, Suha, Lee, Honglak, and Han, Bohyung. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, pp. 2224–2232. IEEE Computer Society, 2017. ISBN 978-1-5386-0457-1.

Houthooft, Rein, Chen, Xi, Chen, Xi, Duan, Yan, Schulman, John, Turck, Filip De, and Abbeel, Pieter. Vime: Variational information maximizing exploration. In *NIPS*, pp. 1109–1117, 2016.

Hurley, K. B. and Oakes, L. M. Experience and distribution of attention: Pet exposure and infants' scanning of animal images. *J Cogn Dev*, 16(1):11–30, Jan 2015.

Hurley, K. B., Kovack-Lesh, K. A., and Oakes, L. M. The influence of pets on infants' processing of cat and dog images. *Infant Behav Dev*, 33(4):619–628, Dec 2010.

Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z., Silver, David, and Kavukcuoglu, Koray. Reinforcement learning with unsupervised auxiliary tasks. *CoRR*, abs/1611.05397, 2016.

Kalchbrenner, Nal, van den Oord, Aron, Simonyan, Karen, Danihelka, Ivo, Vinyals, Oriol, Graves, Alex, and Kavukcuoglu, Koray. Video pixel networks. In *ICML*, volume 70 of *JMLR Workshop and Conference Proceedings*, pp. 1771–1779. JMLR.org, 2017.

Kidd, Celeste, Piantadosi, Steven T., and Aslin, Richard N. The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLOS ONE*, 7(5):1–8, 05 2012. doi: 10.1371/journal.pone.0036399.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Kulkarni, Tejas D., Narasimhan, Karthik, Saeedi, Ardavan, and Tenenbaum, Josh. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *NIPS*, pp. 3675–3683, 2016.

Mitash, Chaitanya, Bekris, Kostas E., and Boularias, Abdeslam. A self-supervised learning system for object detection using physics simulation and multi-view pose estimation. In *IROS*, pp. 545–551. IEEE, 2017. ISBN 978-1-5386-2682-5.

Noroozi, Mehdi and Favaro, Paolo. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV (6)*, volume 9910 of *Lecture Notes in Computer Science*, pp. 69–84. Springer, 2016. ISBN 978-3-319-46465-7.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res*, 37:3311–25, 1997.

Pathak, Deepak, Agrawal, Pulkit, Efros, Alexei A., and Darrell, Trevor. Curiosity-driven exploration by self-supervised prediction. In *ICML*, volume 70 of *JMLR Workshop and Conference Proceedings*, pp. 2778–2787. JMLR.org, 2017.

Popov, Ivaylo, Heess, Nicolas, Lillicrap, Timothy, Hafner, Roland, Barth-Maron, Gabriel, Vecerik, Matej, Lampe, Thomas, Tassa, Yuval, Erez, Tom, and Riedmiller, Martin. Data-efficient deep reinforcement learning for dexterous manipulation. *arXiv preprint arXiv:1704.03073*, 2017.

Saxe, Rebecca and Kanwisher, Nancy. People thinking about thinking people: the role of the temporo-parietal junction in theory of mind. *Neuroimage*, 19(4):1835–1842, 2003.

Schmidhuber, Jürgen. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. Autonomous Mental Development*, 2(3):230–247, 2010.

Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec, and Klimov, Oleg. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Sener, Ozan and Savarese, Silvio. A geometric approach to active learning for convolutional neural networks. *CoRR*, abs/1708.00489, 2017.

Settles, Burr. *Active Learning*, volume 18. Morgan & Claypool Publishers, 2011.

Singh, Satinder P., Lewis, Richard L., Barto, Andrew G., and Sorg, Jonathan. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Trans. Autonomous Mental Development*, 2(2):70–82, 2010.

Sokolov, E.N. *Perception and the conditioned reflex*. Pergamon Press, 1963.

Spyros Gidaris, Praveer Singh, Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018.

Twomey, K. E. and Westermann, G. Curiosity-based learning in infants: a neurocomputational approach. *Dev Sci*, Oct 2017.

Wang, Keze, Zhang, Dongyu, Li, Ya, Zhang, Ruimao, and Lin, Liang. Cost-effective active learning for deep image classification. *IEEE Trans. Circuits Syst. Video Techn.*, 27(12):2591–2600, 2017.

Zhang, Richard, Isola, Phillip, and Efros, Alexei A. Colorful image colorization. In Leibe, Bastian, Matas, Jiri, Sebe, Nicu, and Welling, Max (eds.), *ECCV (3)*, volume 9907 of *Lecture Notes in Computer Science*, pp. 649–666. Springer, 2016. ISBN 978-3-319-46486-2.