



Inria
INVENTEURS DU MONDE NUMÉRIQUE

ENSTA



IP PARIS
ARMINES



IP PARIS

Autonomous Systems and Robotics
Computer Science and System Engineering Laboratory

State Representation Learning for control: an overview

Natalia Díaz Rodríguez, PhD



Special thanks to collaborators: **T Lesort, J-F Goudou, D Filliat**

ABOUT ME



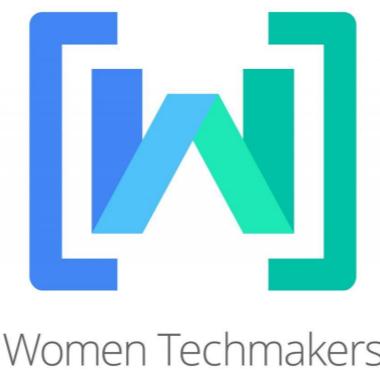
Universidad
de Granada



UNIVERSITY OF CALIFORNIA
SANTA CRUZ



Google
♀

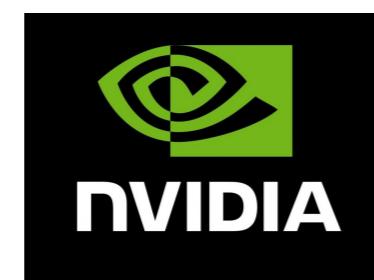


IP PARIS

Autonomous Systems and Robotics
Computer Science and System Engineering Laboratory



HEIDELBERG
LAUREATE
FORUM



PHILIPS



STITCH FIX™

asteria



2



LIFELONG LEARNING AND DEVELOPMENT IN ROBOTS AND HUMANS



The [Flowers project-team](#), at Inria and Ensta ParisTech, studies mechanisms that can allow robots and humans to acquire autonomously and cumulatively repertoires of novel skills over extended periods of time.

This includes mechanisms for learning by self-exploration, as well as learning through interaction with peers, for the acquisition of both sensorimotor and social skills. Sensorimotor skills include locomotion, affordance learning, active manipulation. Interactive skills include grounded language use and understanding, adaptive interaction protocols, and human-robot collaboration.

Our approach is organized along two strands of research:

Artificial intelligence: constructing machines and robots, inspired by animal cognitive development, and capable of lifelong development, adaptation and interaction with the physical and social world.

Cognitive Science: Elaborating computer and robotic models as tools for understanding developmental processes in humans.

Our project-team, headed by [Pierre-Yves Oudeyer](#) (Inria) and co-started with [David Filliat](#) (Ensta ParisTech [Cognitive Robotics Group](#)), focuses in particular on the study of developmental mechanisms that guide efficient open-ended learning of novel skills in large real world environments. In particular, we study:

- **Intrinsically motivated multitask learning and exploration, information seeking and active learning, including artificial curiosity;**
- **Social learning**, e.g. learning by imitation or demonstration, which implies both issues related to machine learning and human-robot interaction;
- Mechanisms for learning to **sequence and compose actions to reach goals**, especially within the framework of reinforcement learning;
- The role of **embodiment**, in particular through the concept of morphological computation, as well as the structure of motor primitives/muscle synergies that can leverage the properties of morphology and physics;
- **Maturational constraints** which can allow the progressive release of novel sensorimotor degrees of freedom to be explored;

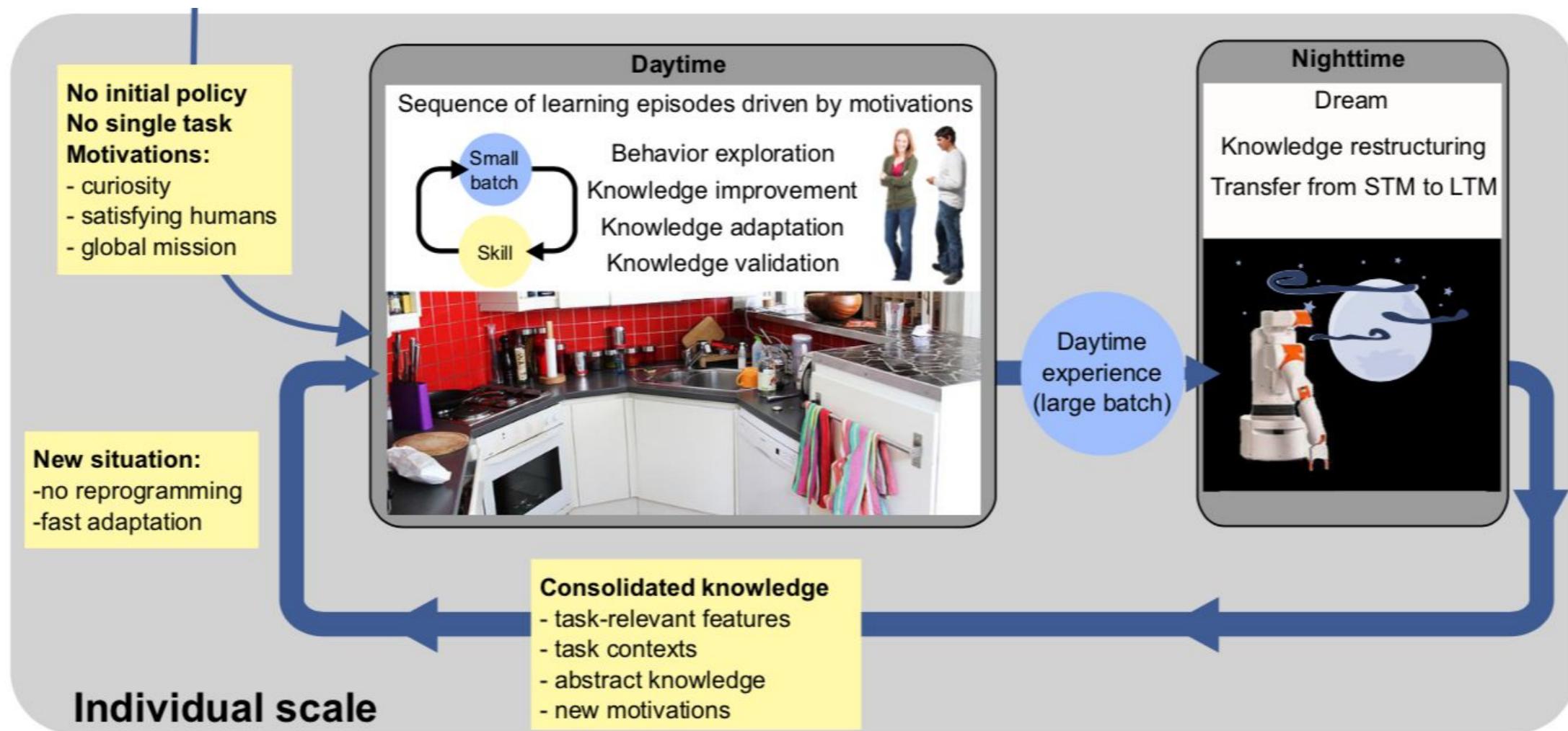
DREAM Project:

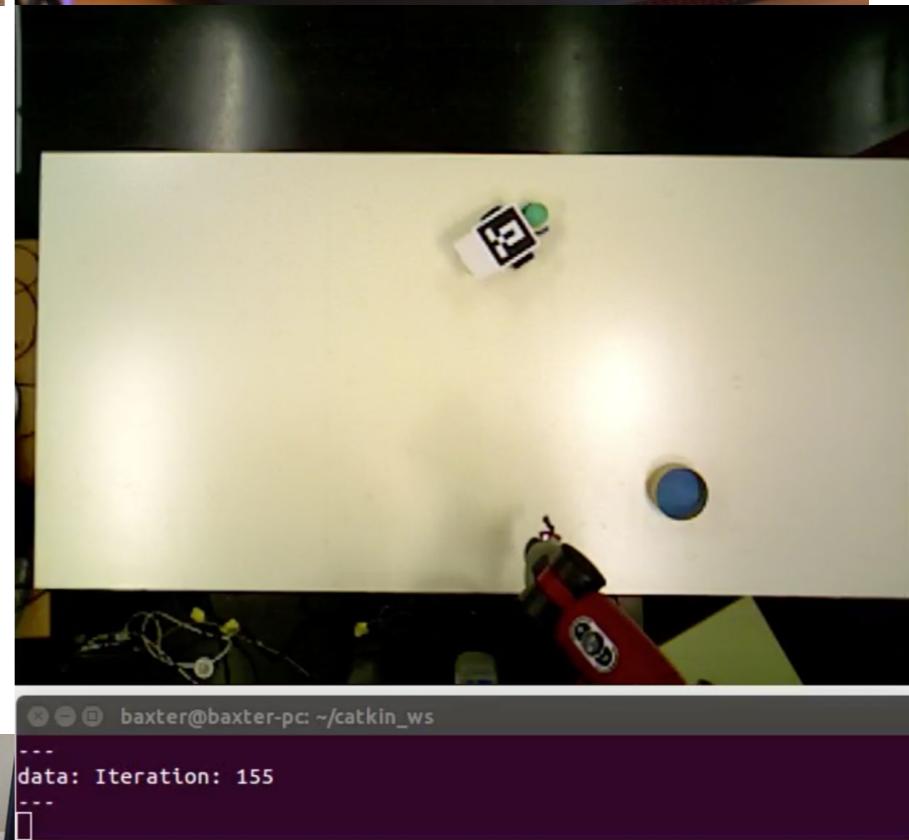


Rethink robotics' Baxter



- Learning and adapting in open-ended domains
 - Knowledge compression, consolidation, and transfer
 - Artificial curiosity and representation redescription
- **Day:** getting data and learning tasks. **Night:** Finding better representations





This presentation

Outline

Abstract

Keywords

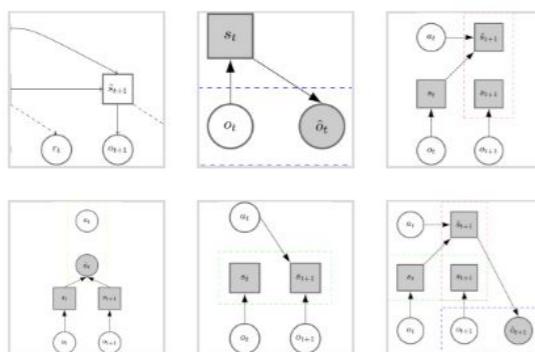
1. Introduction
2. Formalism and definitions
3. Learning objectives
4. Building blocks of state representation learning
5. Discussion and future trends
6. Conclusion

Acknowledgments

References

Show full outline 

Figures (7)



Show all figures 

Tables (3)

- Table 1
- Table 2
- Table 3



Neural Networks

Volume 108, December 2018, Pages 379-392



Review

State representation learning for control: An overview

Timothée Lesort ^{a, b} , Natalia Díaz-Rodríguez ^b , Jean-François Goudou ^a , David Filliat ^b 

 Show more

<https://doi.org/10.1016/j.neunet.2018.07.006>

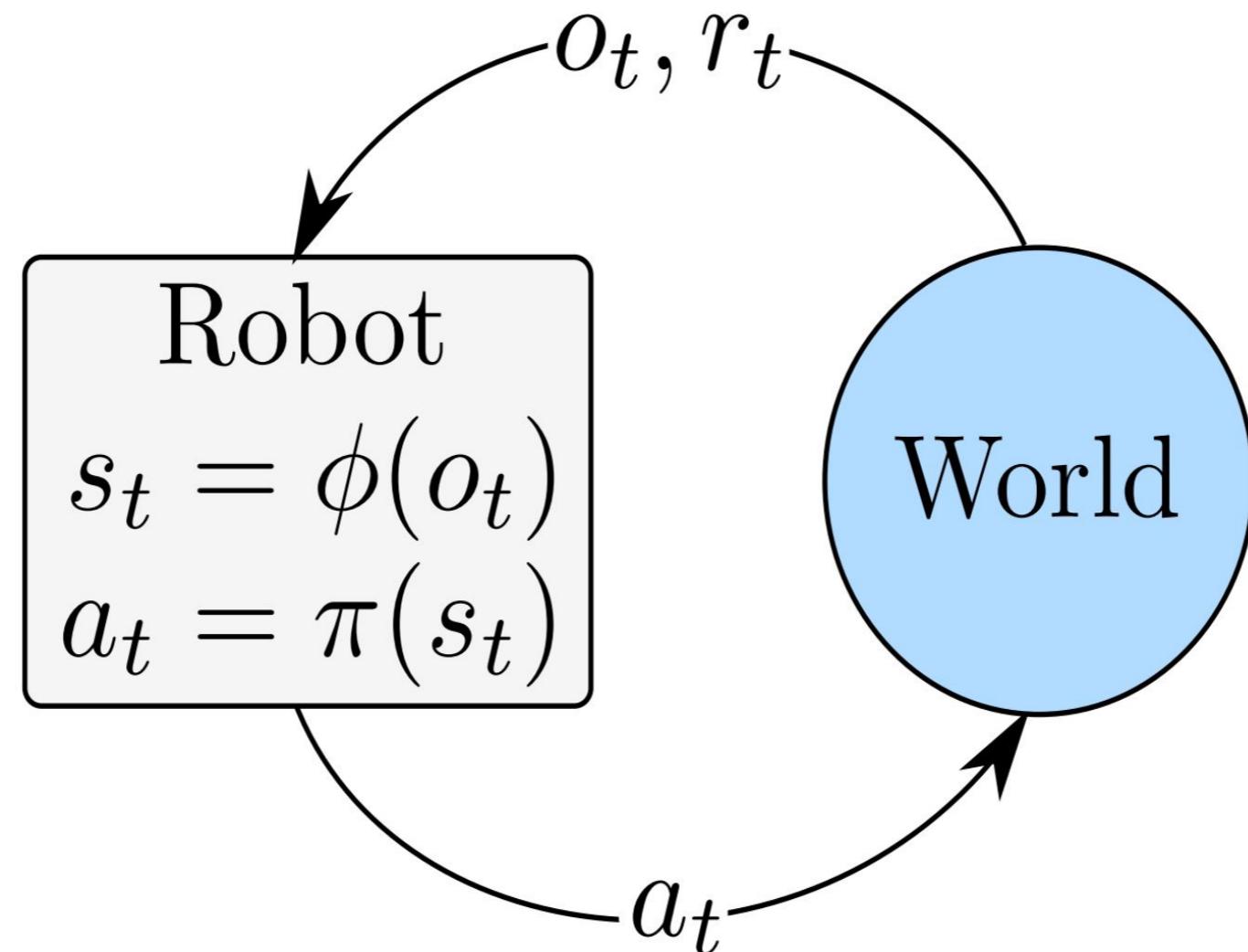
[Get rights and content](#)

Abstract

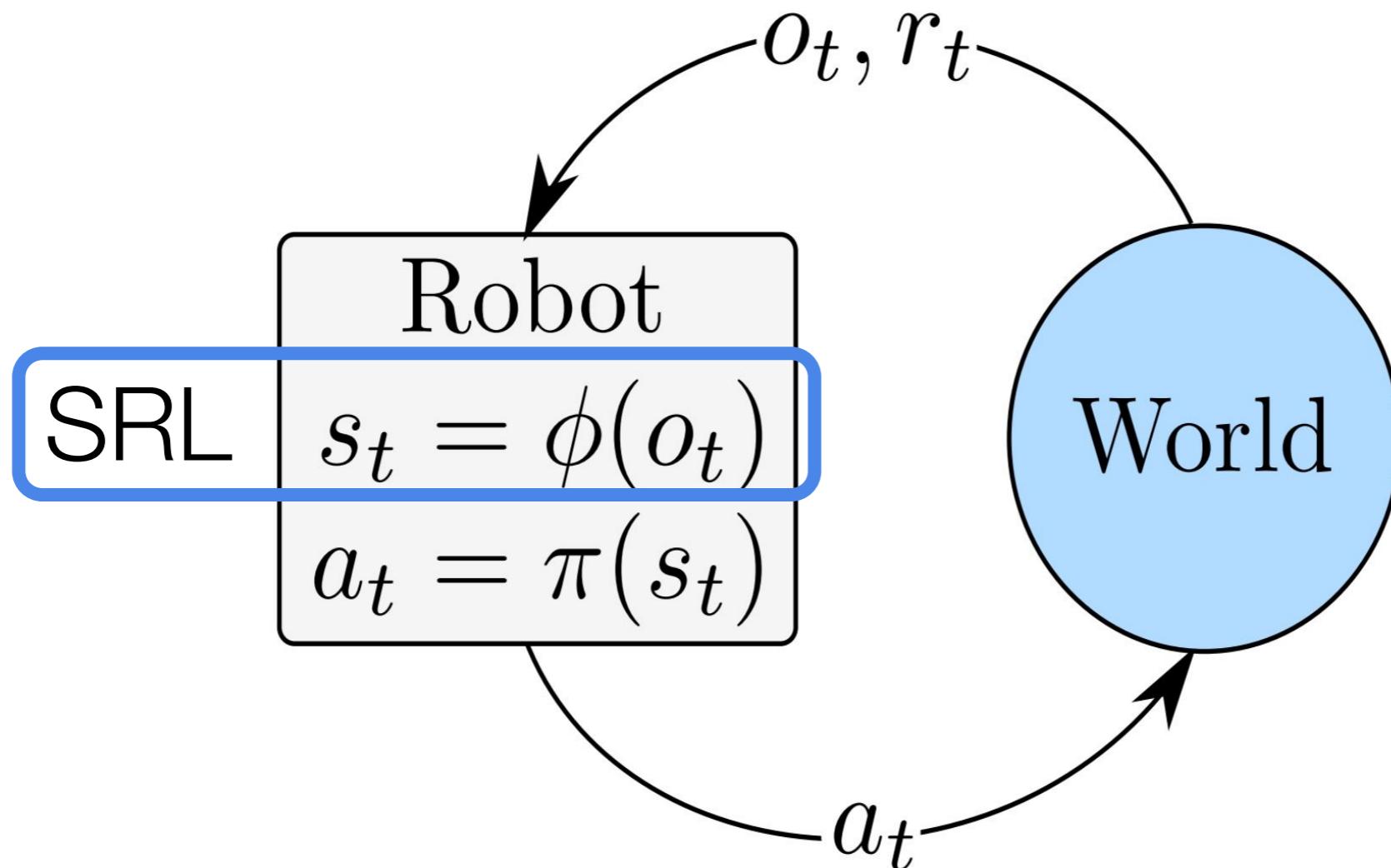
Representation **learning algorithms** are designed to learn abstract features that characterize data. State representation learning (SRL) focuses on a particular kind of representation learning where learned features are in low dimension, evolve through time, and are influenced by actions of an agent. The representation is learned to capture the variation in the environment generated by the agent's actions; this kind of representation is particularly suitable for **robotics** and control scenarios. In particular, the low **dimension characteristic** of the representation helps to overcome the curse of **dimensionality**, provides easier interpretation and utilization by humans and can help improve performance and speed in policy learning algorithms such as **reinforcement learning**.

This survey aims at covering the state-of-the-art on state representation learning in the most recent years. It reviews different SRL methods that **involve interaction** with the environment, their implementations and their applications in robotics control tasks (simulated or real). In particular, it highlights how generic learning objectives are differently exploited in the reviewed algorithms. Finally, it discusses evaluation methods to assess the representation learned and summarizes current and future lines of research.

State Representation Learning (SRL) in RL context



State Representation Learning (SRL) in RL context



Why learning states ?



- Often, robot controllers require simple, ‘high-level’ inputs (the ‘state’ of the world)
 - E.g., *grasping*: object position; *driving*: road direction
- High dimensional visual input perception requires filtering to get this relevant information (often hand-crafted)
 - @DREAM: Learning state = Discarding irrelevant knowledge (e.g., during ‘off’ or ‘sleep time’).
- Controllers are easier to train in such lower dimension
 - Less dependent on the environment
 - Help transfer learning

Learning states vs learning representations



- Learning a state:
 - A particular case of representation learning
 - Entails a control/robotics context where actions are possible
 - Can be an objective by itself, but is often present in more general approaches
 - E.g., as an auxiliary task in RL
- Often we look for interpretable info./ info. with a *physical* meaning

Learning states vs learning representations



- Focus in this survey:
 - Learning objectives relevant for SRL
 - Put in perspective common factors between very different use cases

First, what is a good state?



The Markovian view: [Böhmer'15]

A good state representation is a representation that is:

1. **Markovian**, i.e. it summarizes all the necessary information to be able to choose an action within the policy, by looking *only* at the current state.
2. Able to:
 - a. Represent the true value of the current state well enough for **policy improvement**.
 - b. **Generalize** the learned value-function to unseen states with similar futures.
3. **Compact**: Low dimensional for efficient estimation.

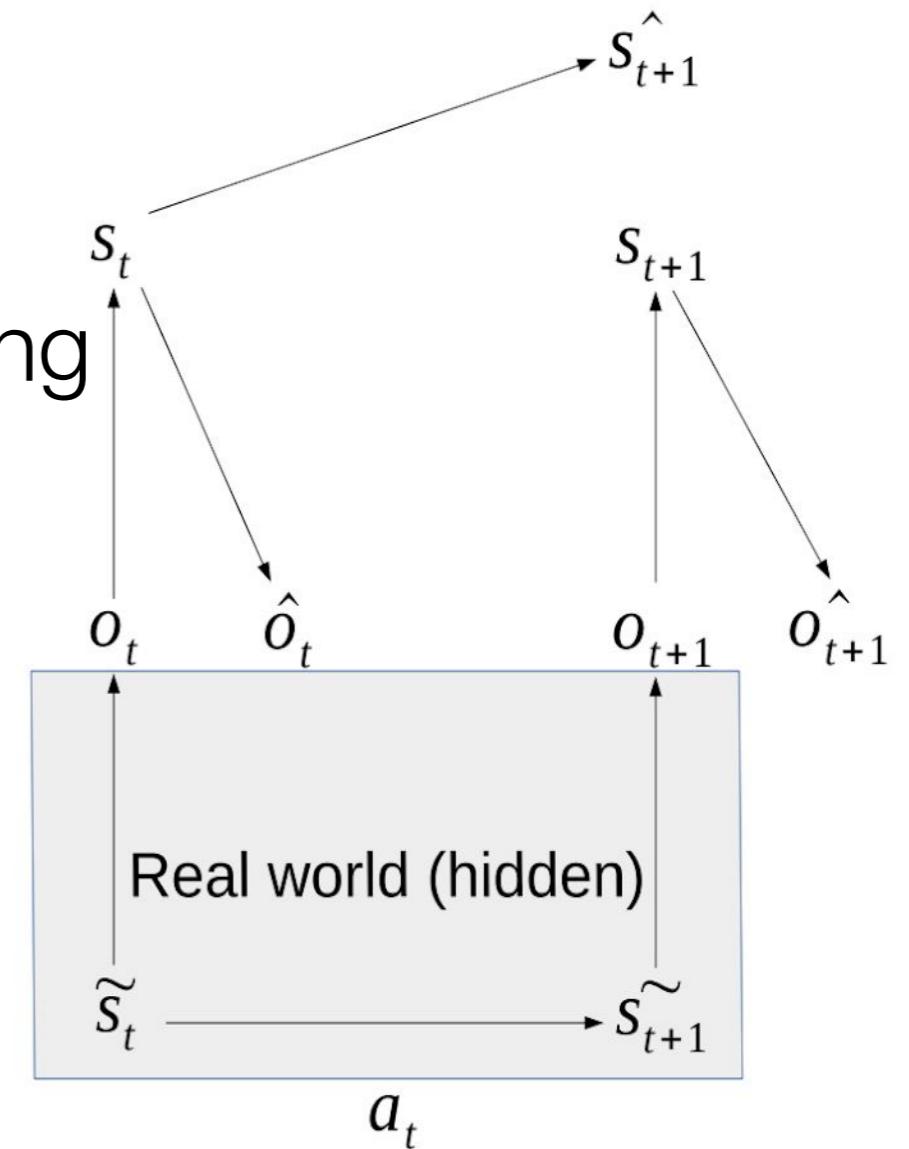
Further, a good state representation should:

- Clearly **disentangle** the different factors of variation with different semantics
- **Sufficient**
- As **efficient** as possible (easy to work with, e.g., factorizing the data-generating factors)
- **Minimal** (from all possible representations, the most efficient one).

=> Exclude irrelevant information to encourage a lower dimensionality

Learning states: challenges

- Learning without direct supervision
 - True state never available
- Solution: Use ‘self-supervised’ learning
 - RL framework
 - Exploit observations, actions, rewards
- Various complementary approaches
 - Reconstruct observation
 - Learn forward/inverse model
 -



What can SRL be used for? e.g., in computer vision for robotics control

Our goal is to learn a relevant state from images, actions and rewards. We train a neural network in unsupervised manner using prior knowledge about the physical world in form of robotic priors.

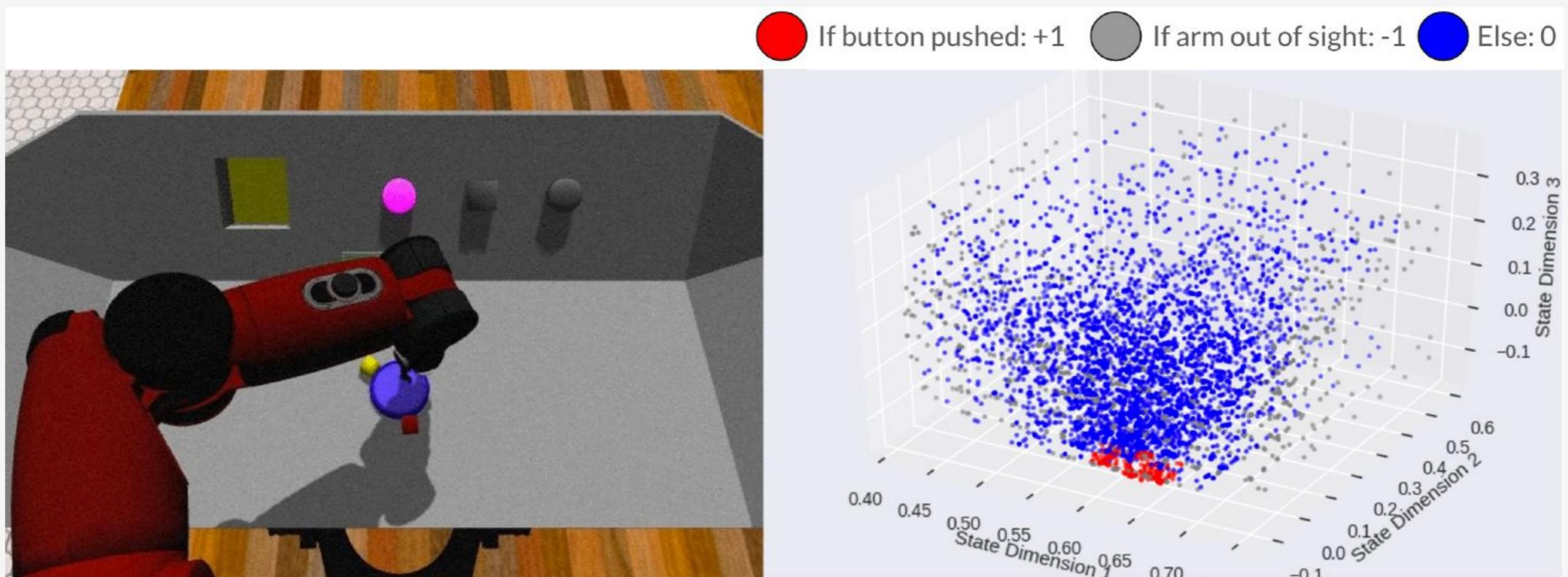
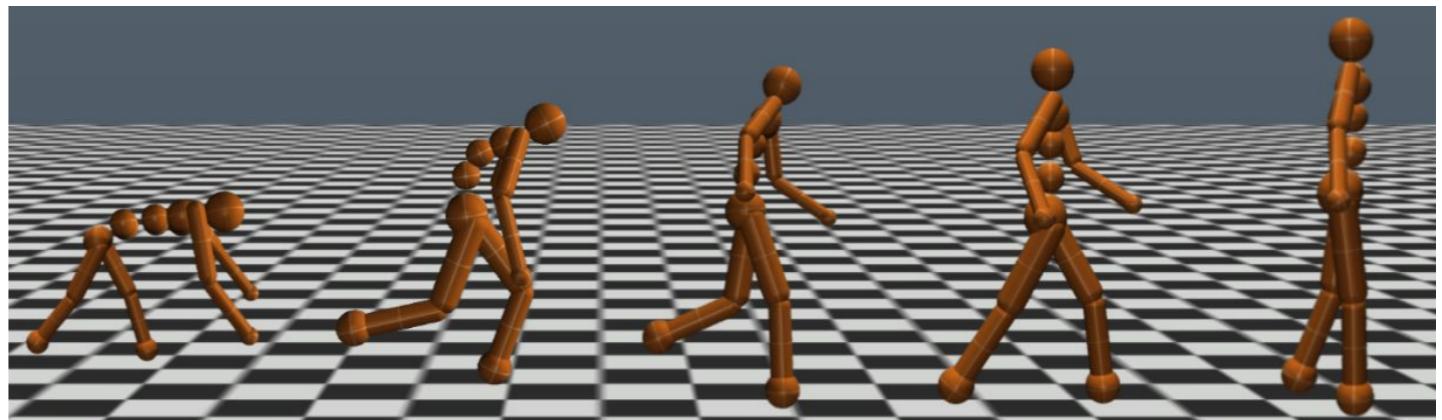


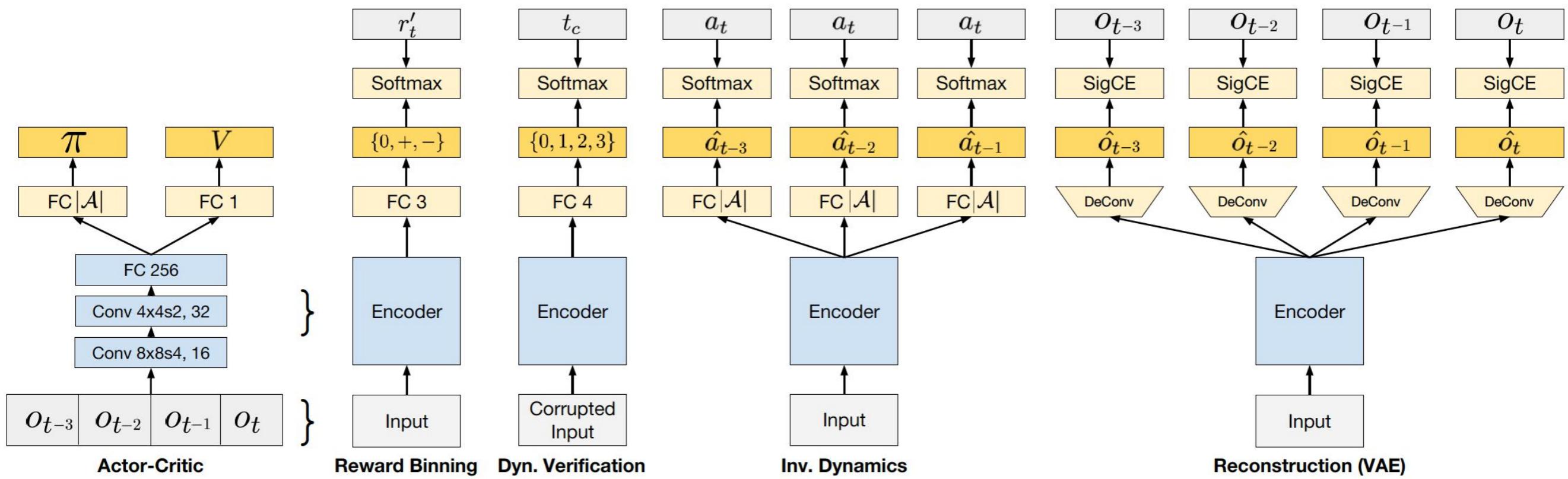
Figure 1: Left: Baxter's camera view for Static-Button-Distractors dataset 2. Right: Baxter's left hand ground truth position and its coded reward

Applications of SRL

- RL and self-supervision
- Transfer learning
- Multimodal observation learning (e.g., autonomous vehicles)
- Multi-objective optimization
- Evolutionary strategies

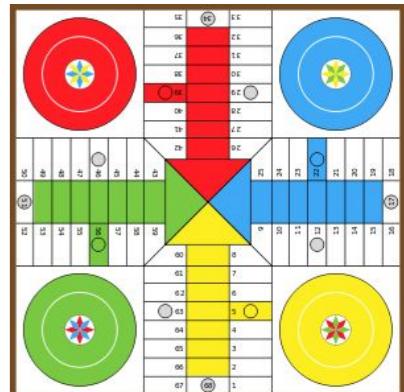
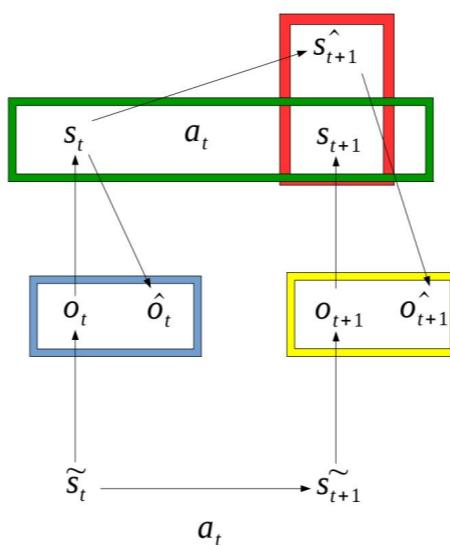
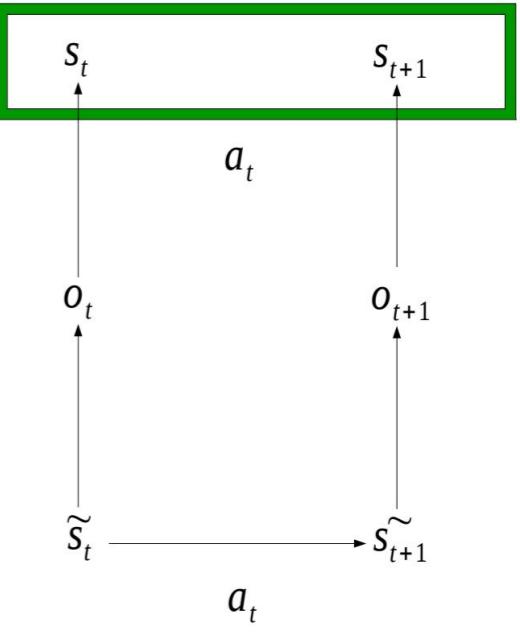
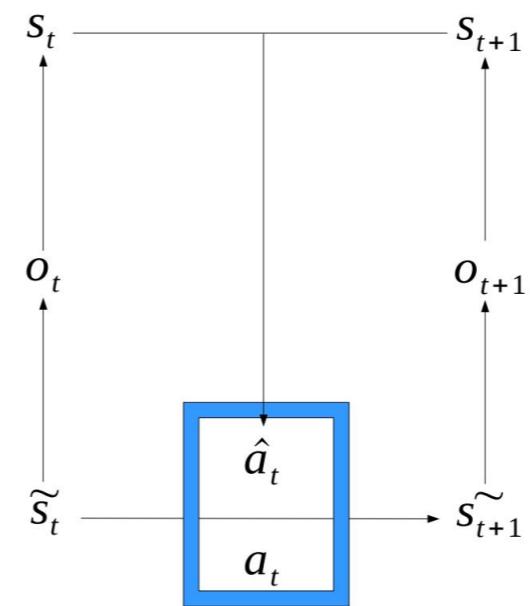
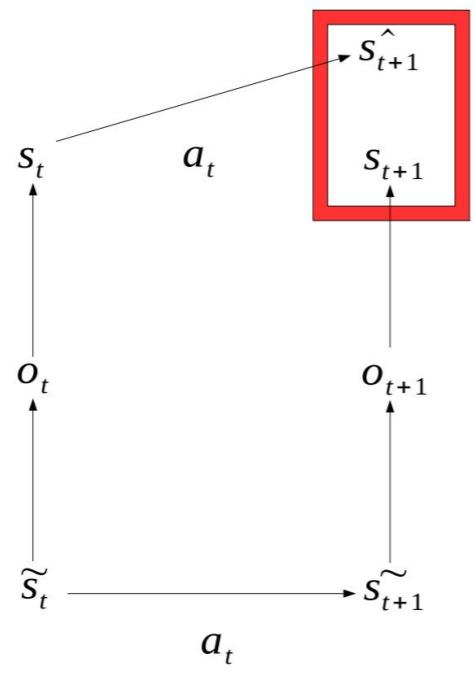
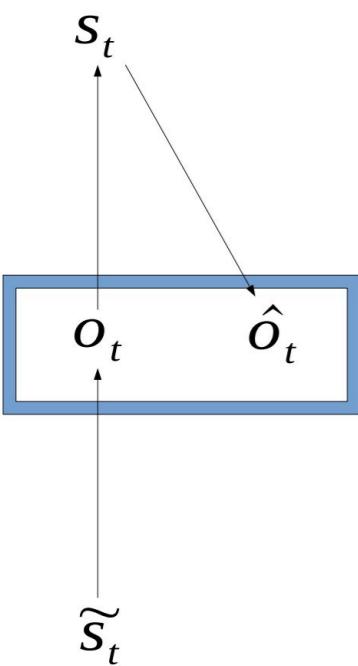


Architectures for RL & Self-supervision



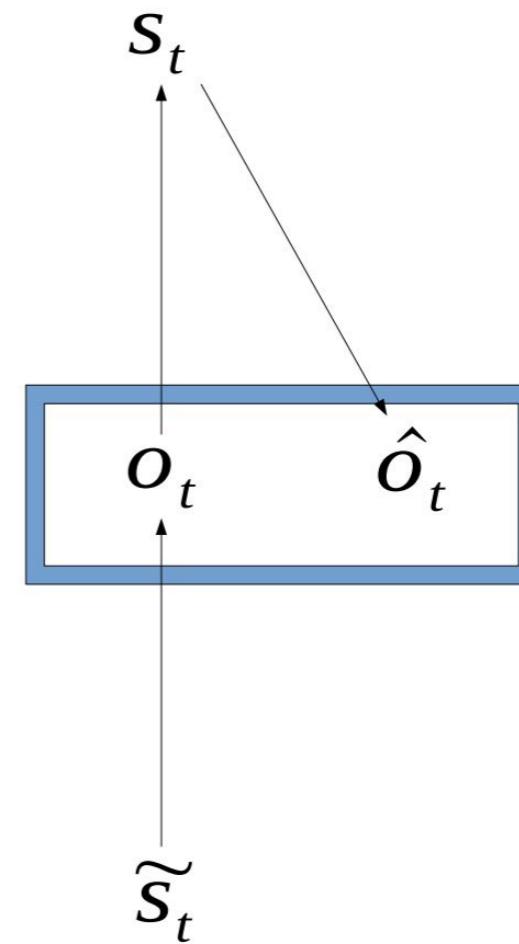
[Shelhamer'17 Loss is its own reward: Self-supervision for reinforcement learning.
arXiv preprint arXiv:1612.07307.]

Overview of existing approaches



Models that reconstruct the observation

- Train a state that is sufficient for reconstructing the input observation
 - AE, DAE, VAE
 - GANs
- Downside: sensitive to irrelevant variations



(a) Auto-Encoder

$$s_t = \phi(o_t; \theta_\phi) \tag{1}$$

$$\hat{o}_t = \phi^{-1}(s_t; \theta_{\phi^{-1}}) \tag{2}$$

AE example: Deep Spatial Autoencoders

- Learns a representation that corresponds to objects coords. (image space):
feature encoder $h_{\text{enc}}(I_t) = \mathbf{f}_t$
- Torque control & 2nd-order dynamical system: require features and their velocities $\dot{\mathbf{f}}_t$
full state $\mathbf{x}_t = [\tilde{\mathbf{x}}_t; \tilde{\mathbf{f}}_t; \dot{\mathbf{f}}_t]$

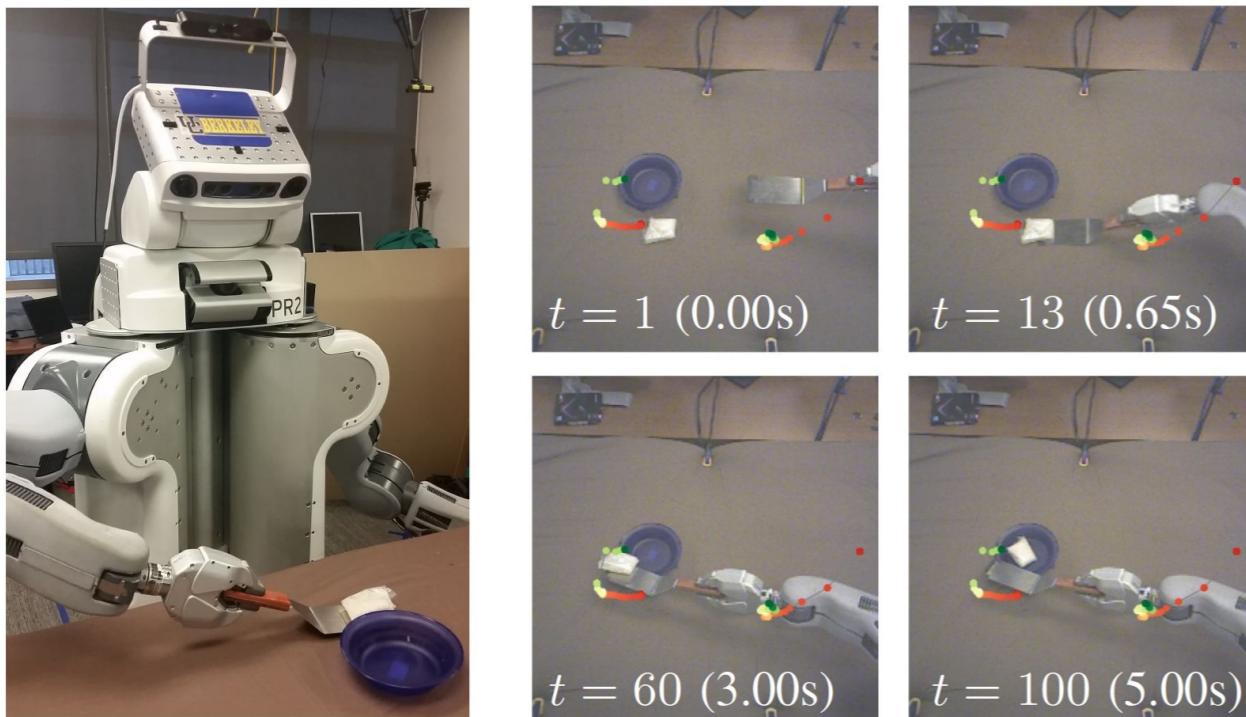


Fig. 1: PR2 learning to scoop a bag of rice into a bowl with a spatula (left) using a learned visual state representation (right).

AE example: Deep Spatial Autoencoders

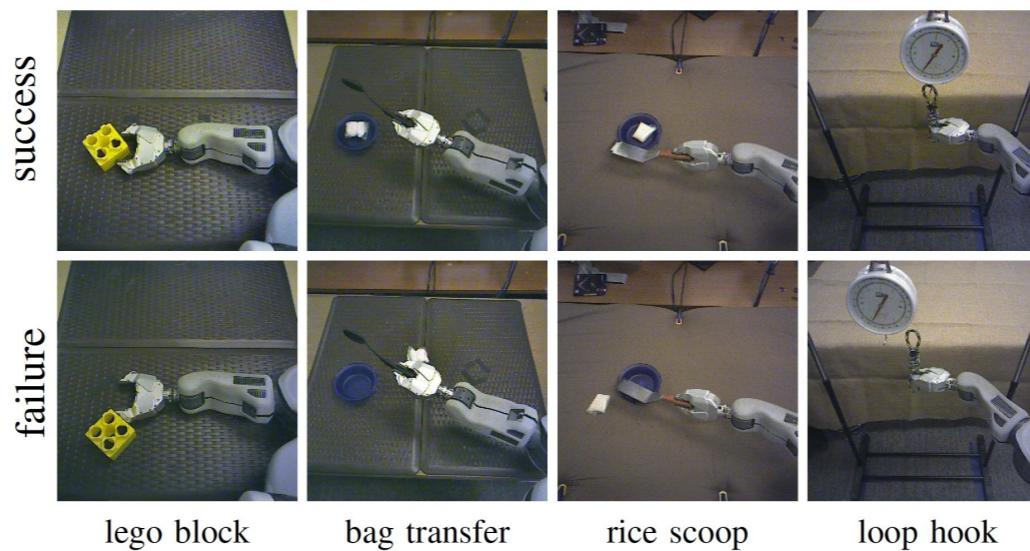
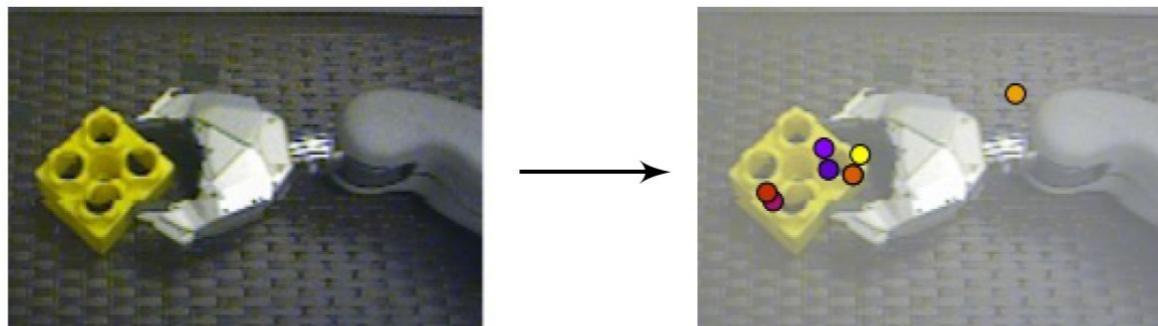
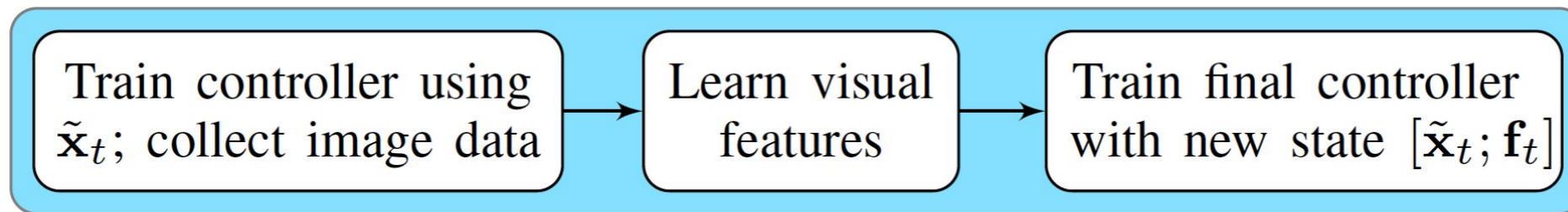
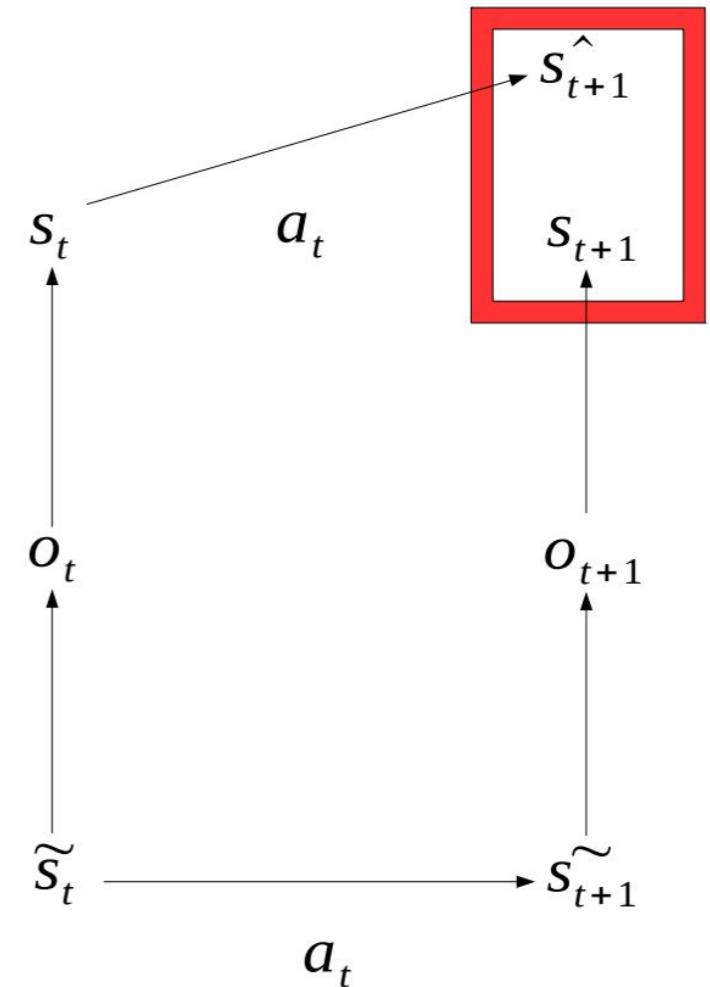


Fig. 5: Illustrations of the tasks in our experiments, as viewed by the robot's camera, showing examples of successful completion by our controllers and a typical failure by the controller without vision.

Forward models:

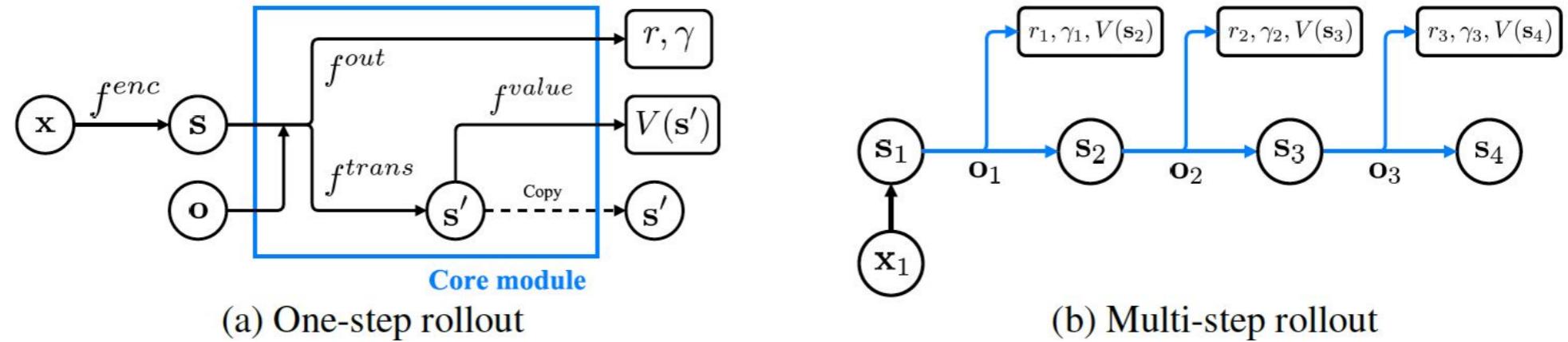
- Find state from which it is easy to predict next state
 - Impose constraints (e.g., simple linear model)
 - Naturally discard irrelevant features
- May be useful in, e.g. model based RL



(b) Forward Model

$$\hat{s}_{t+1} = f(s_t, a_t; \theta_{fwd})$$

Forward model example: Value Prediction Network [Oh'17]



The VPN consists of the following modules parameterized by $\theta = \{\theta^{enc}, \theta^{value}, \theta^{out}, \theta^{trans}\}$:

Encoding $f_\theta^{enc} : \mathbf{x} \mapsto \mathbf{s}$

Outcome $f_\theta^{out} : \mathbf{s}, \mathbf{o} \mapsto r, \gamma$

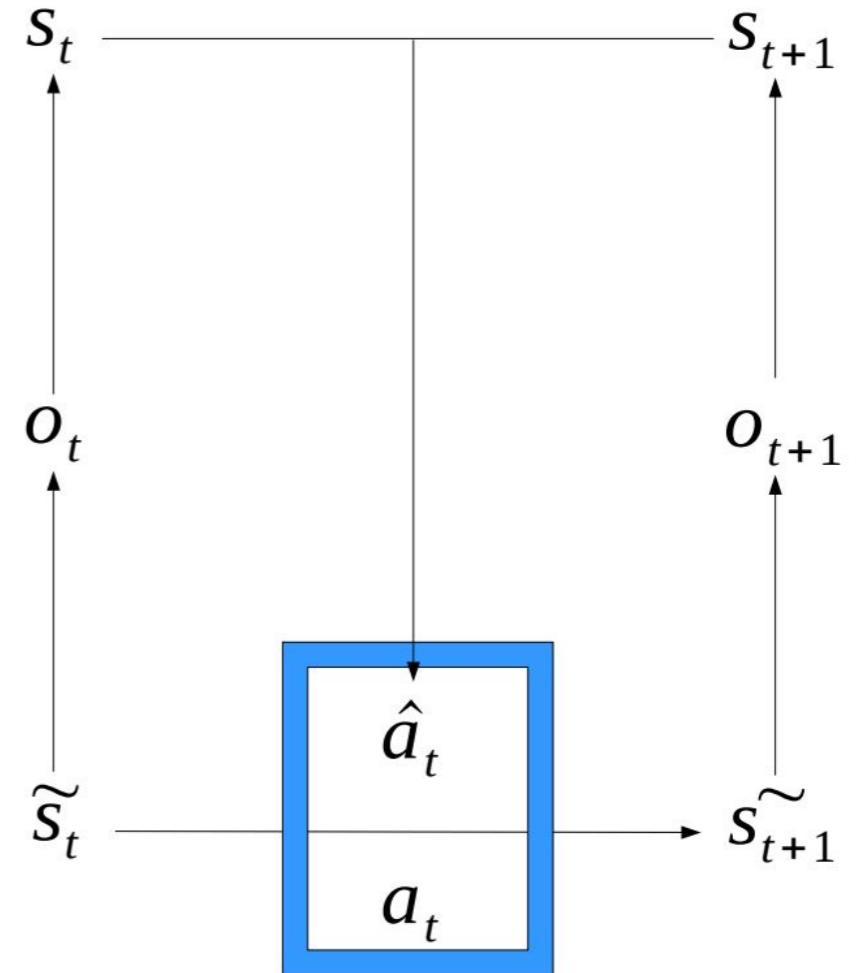
Value $f_\theta^{value} : \mathbf{s} \mapsto V_\theta(\mathbf{s})$

Transition $f_\theta^{trans} : \mathbf{s}, \mathbf{o} \mapsto \mathbf{s}'$

- SRL happens in the **Transition** (*fwd*) model
- Learns the dynamics of the reward values to perform planning (based on **options**)

Inverse models

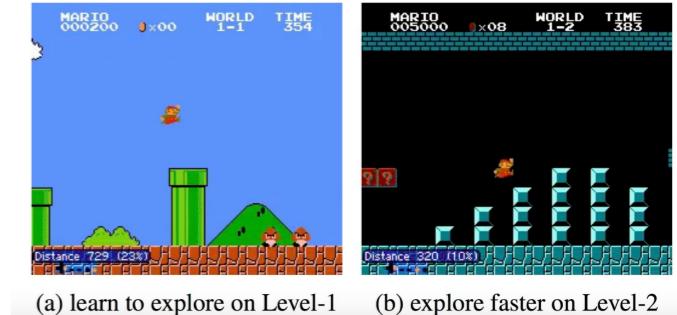
- Train a state:
 - Sufficient to find actions from 2 observations
 - Useful for a direct control model
- Impose constraints on the model, e.g.:
 - Simple linear model
 - Focus on states that can be controlled



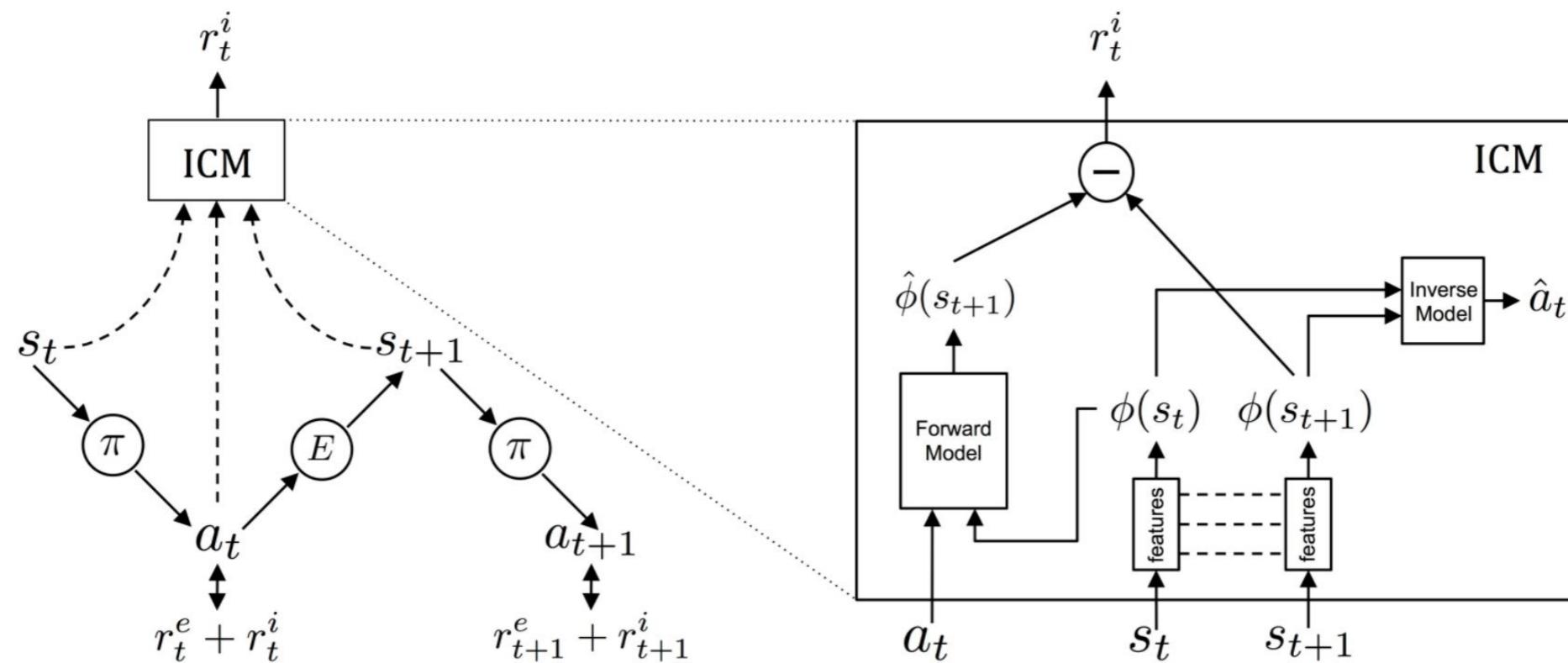
(c) Inverse Model

$$\hat{a}_t = g(s_t, s_{t+1}; \theta_{inv})$$

The role of SRL in the intrinsic curiosity model

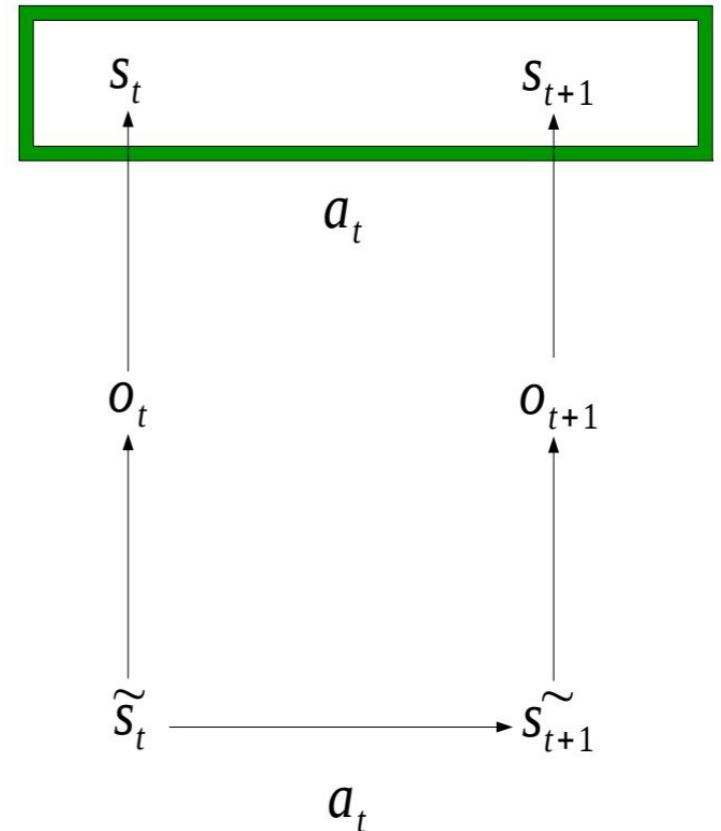


- The *fwd* model encodes constraints & learns the state space
- Its prediction error is the intrinsic (i.e., *curiosity*) reward that encourages space exploration



Priors models

- Encode high-level constraints on the states:
 - Temporal continuity
 - Controllability
 - Inertia, etc.
- May exploit rewards



(d) Model with prior

$$Loss = \mathcal{L}_{prior}(s_1, \dots, s_n; \theta_\phi | c) \quad (5)$$

Priors models

In case you wonder...

we do NOT mean PRIORS in the *bayesian* prior probability sense. Rather:

a priori

Something that can be known
without experience or sense
data.

If Socrates has more wine than Plato and Plato has more wine than Aristotle, then Socrates has more wine than Aristotle.

Five is a prime number.

Brothers are male siblings.

Carneades.org

Robotic Priors

[Jonschkowski et. al. 2015]

Use a *priori* knowledge to learn representations more relevant to the task

Robotic Priors [1] provide the model with basic knowledge about the environment dynamical features. Each prior is encoded as a loss function:

- ▶ **Temporal coherence Prior:** *Two states close to each other in time are also close to each other in the state representation space.*

$$L_{Temp}(D, \hat{\phi}) = E[\| \Delta \hat{s}_t \|^2], \quad (1)$$

- ▶ **Proportionality Prior:** *Two identical actions should result in two proportional magnitude state variations.*

$$L_{Prop}(D, \hat{\phi}) = E[(\| \Delta \hat{s}_{t_2} \| - \| \Delta \hat{s}_{t_1} \|)^2 | a_{t_1} = a_{t_2}], \quad (2)$$

- ▶ **Repeatability Prior:** *Two identical actions applied at similar states should provide similar state variations, not only in magnitude but also in direction.*

$$L_{Rep}(D, \hat{\phi}) = E[e^{-\|\hat{s}_{t_2} - \hat{s}_{t_1}\|^2} \| \Delta \hat{s}_{t_2} - \Delta \hat{s}_{t_1} \|^2 | a_{t_1} = a_{t_2}], \quad (3)$$

- ▶ **Causality Prior:** *If two states on which the same action is applied give two different rewards, they should not be close to each other in the state representation space.*

$$L_{Caus}(D, \hat{\phi}) = E[e^{-\|\hat{s}_{t_2} - \hat{s}_{t_1}\|^2} | a_{t_1} = a_{t_2}, r_{t_1+1} \neq r_{t_2+1}], \quad (4)$$

Robotic Priors: Position Velocity Encoders

(PVE) [Jonschkowski et. al. 2017]

PVE encode an observation into a position state and velocity estimation.

$$\mathbf{s}_t^{(v)} = \alpha(\mathbf{s}_t^{(p)} - \mathbf{s}_{t-1}^{(p)}), \quad \mathbf{s}_t^{(a)} = \mathbf{s}_t^{(v)} - \mathbf{s}_{t-1}^{(v)}. \quad \mathbf{s}_t = \begin{bmatrix} \mathbf{s}_t^{(p)} \\ \mathbf{s}_t^{(v)} \end{bmatrix}.$$

- **Variation:** Positions of relevant things vary: their representations should do as well.
- **Slowness:** Positions change slowly (velocities are slow)
- **Inertia:** Velocities change slowly (slowness applied to velocities).
- **Conservation:** Velocity magnitudes change slowly (law of conservation of energy: similar amount of energy (*movement*) in consecutive time steps)
- **Controllability:** Controllable things (whose accelerations correlate with the actions of the robot) are relevant.

Robotic Priors

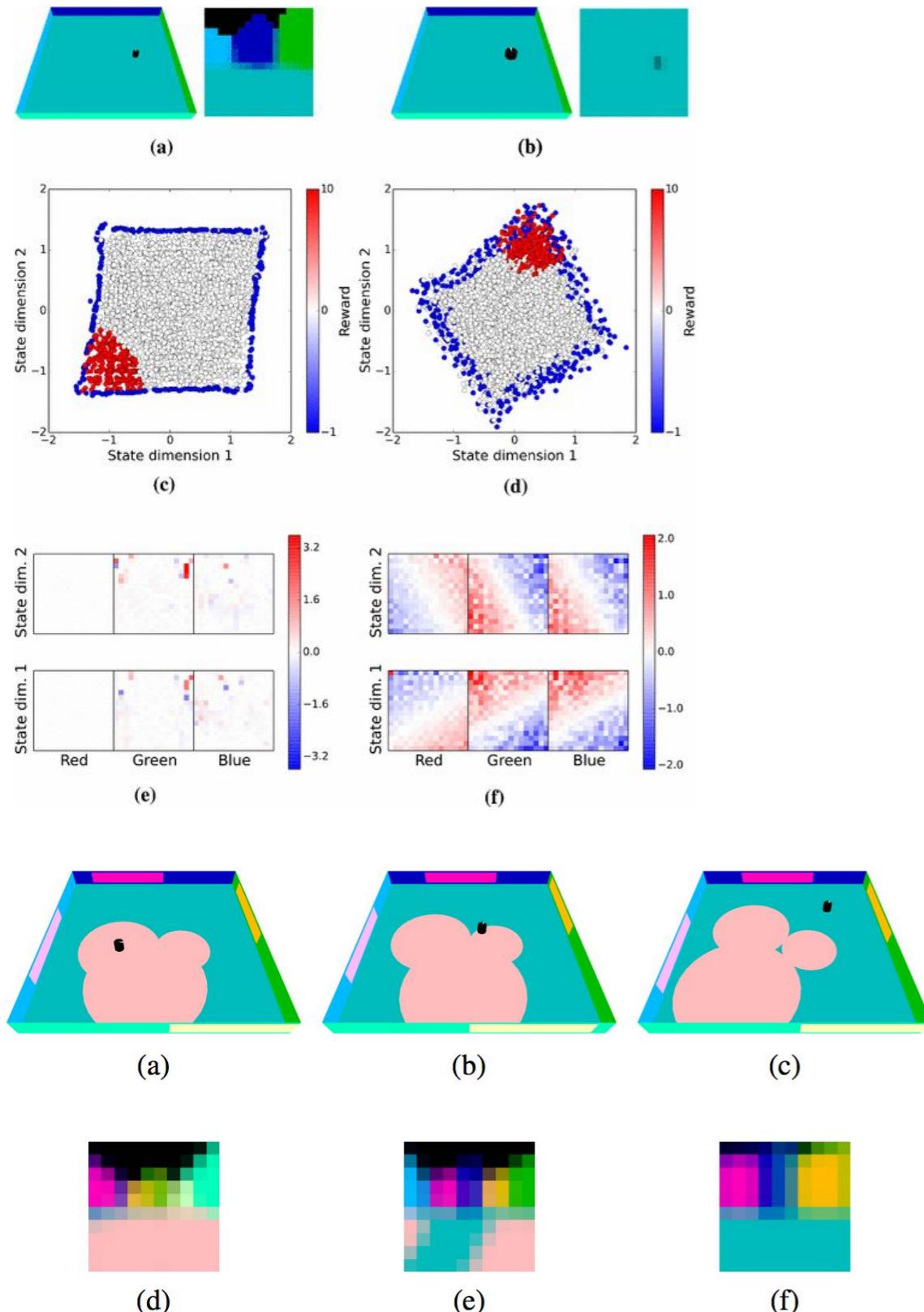


Fig. 7. Extended navigation task. (a-c) show the robot moving to the upper right corner while the distractors move randomly. (d-f) show the respective observations (note how they are influenced by the distractors).

[Jonschkowski et. al. 2015]

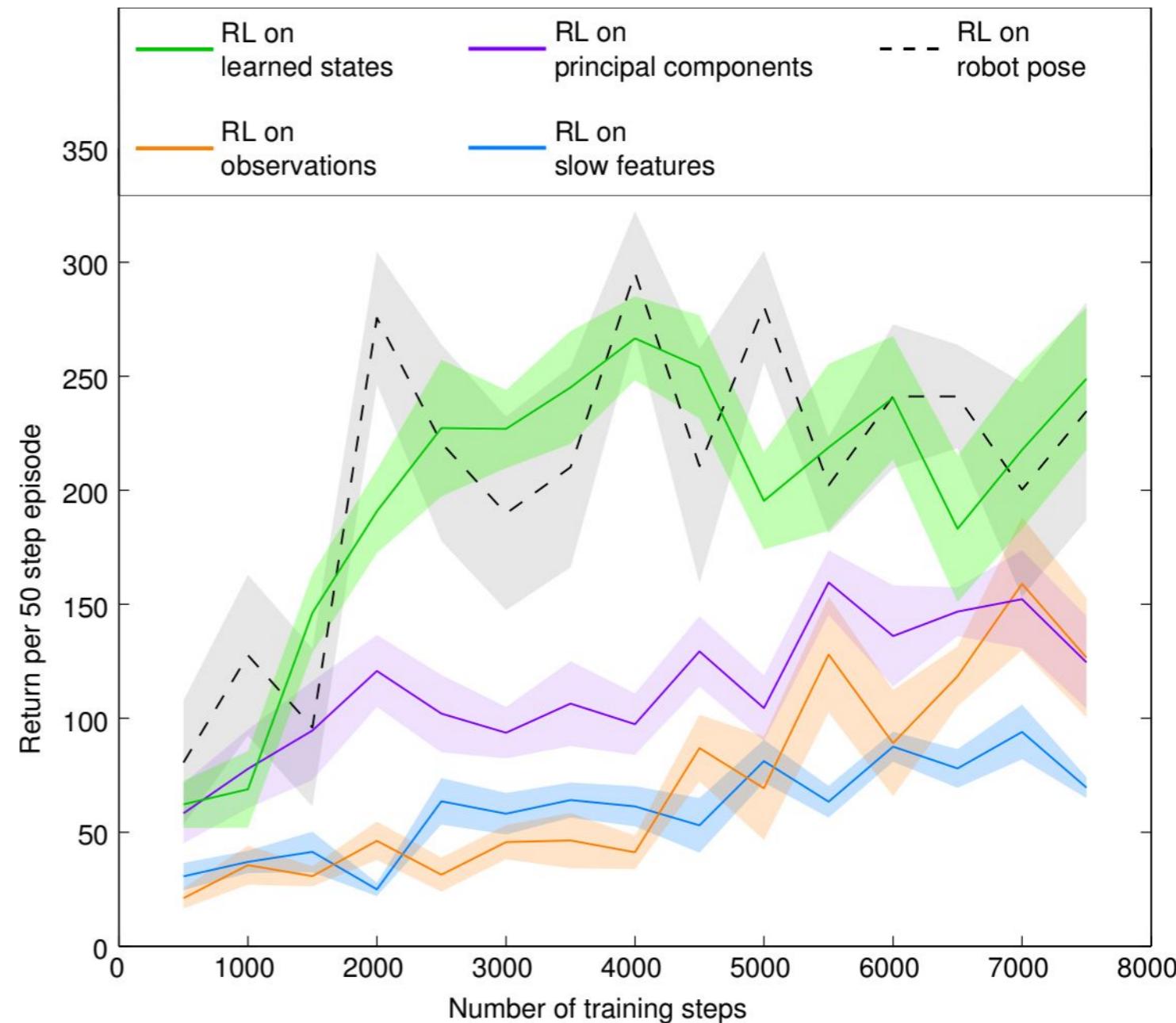


Fig. 8. Reinforcement learning performance for different state representations. Lines show means, surfaces display their standard errors.

Other objective functions

Selectivity: [Thomas'17]

- *A good representation should be a disentangled representation of variation factors.*
- If actions are known to be independent, it measures the independence among variations of the representation under each action.

$$\mathcal{L}_{sel}(D, \phi, k) = \mathbb{E} \left[\frac{\| s_{t+1}^{(k)} - s_t^{(k)} \|}{\sum_{k'} \| s_{t+1}^{(k')} - s_t^{(k')} \|} \mid s_{t+1} \sim P_{s_t, s_{t+1}}^a \right] \quad (16)$$

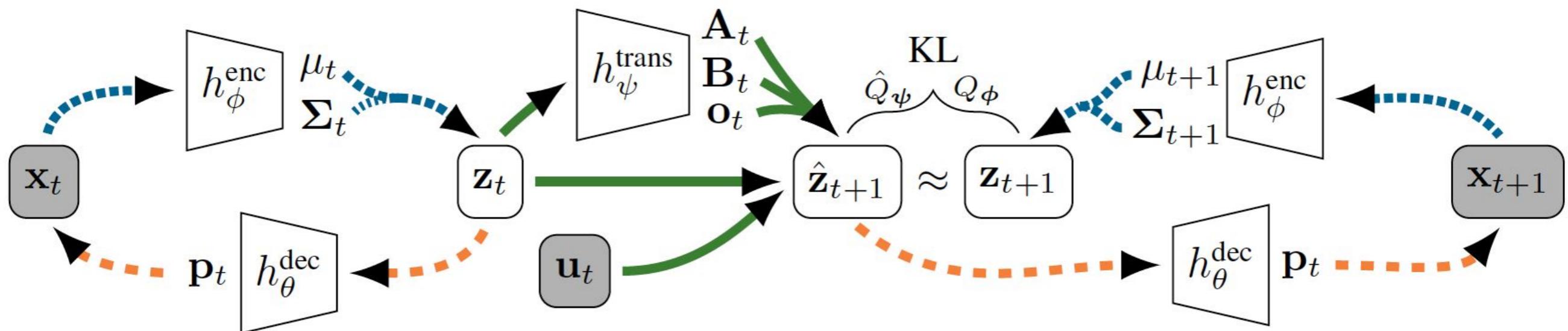
Multi-criteria objective functions: Embed to Control (E2C)

[Watter'18]

- A model to learn the mean and supplementary params for the variance of the distrib. (6)
 - W, U, V fixed or learned
 - Allows using KL-divergence to train a fwd model.

$$\hat{s}_{t+1} = W * \hat{s}_t + U * a_t + V \quad (6)$$

$$\hat{s}_{t+1} \sim \mathcal{N}(\mu = W * \hat{s}_t + U * a_t + V, \sigma) \quad (7)$$



SRL models - summary

Model	Actions/Next state constraints	Forward model	Inverse model	Reconstruct observation	Predicts next observation	Uses rewards
AE [Mattner et al., 2012]	no	no	no	yes	no	no
Priors [Jonschkowski and Brock, 2015]	yes	no	no	no	no	yes
PVE [Jonschkowski et al., 2017]	yes	no	no	no	no	no
E2C [Watter et al., 2015]	yes	yes	no	yes	yes	no
ML-DDPG [Munk et al., 2016]	yes	yes	no	no	no	yes
VAE/DAE [van Hoof et al., 2016]	yes	yes	no	yes	yes	no
AE [Finn et al., 2015]	yes	no	no	yes	no	no
DVBF [Karl et al., 2016]	yes	yes	no	yes	yes	no
[Goroshin et al., 2015]	yes	yes	no	yes	yes	no
ICM [Pathak et al., 2017]	yes	yes	yes	no	no	no

SRL models - summary

Model	Actions/Next state constraints	Forward model	Inverse model	Reconstruct observation	Predicts next observation	Uses rewards
[Shelhamer et al., 2017]	yes	no	yes	no	no	no
VPN [Oh et al., 2017]	no	yes	no	no	no	yes
DDM [Assael et al., 2015]	yes	yes	no	yes	yes	no
AE [Wahlström et al., 2015]	yes	yes	no	yes	yes	no
[Thomas et al., 2017]	yes	no	no	yes	no	no
PCA [Curran et al., 2015]	no	no	no	yes	no	no
PCA [Curran et al., 2016]	no	no	no	yes	no	no
rwPCA [Parisi, 2017]	no	no	no		no	yes
[Magrans de Abril and Kanai, 2018]	yes	yes	no	no	no	yes
InfoGAN [Chen et al., 2016]	no	no	no	yes	no	no
BiGAN [Donahue et al., 2016]	no	no	no	yes	no	no
[Duan, 2017]	yes	yes	yes	yes	yes	no

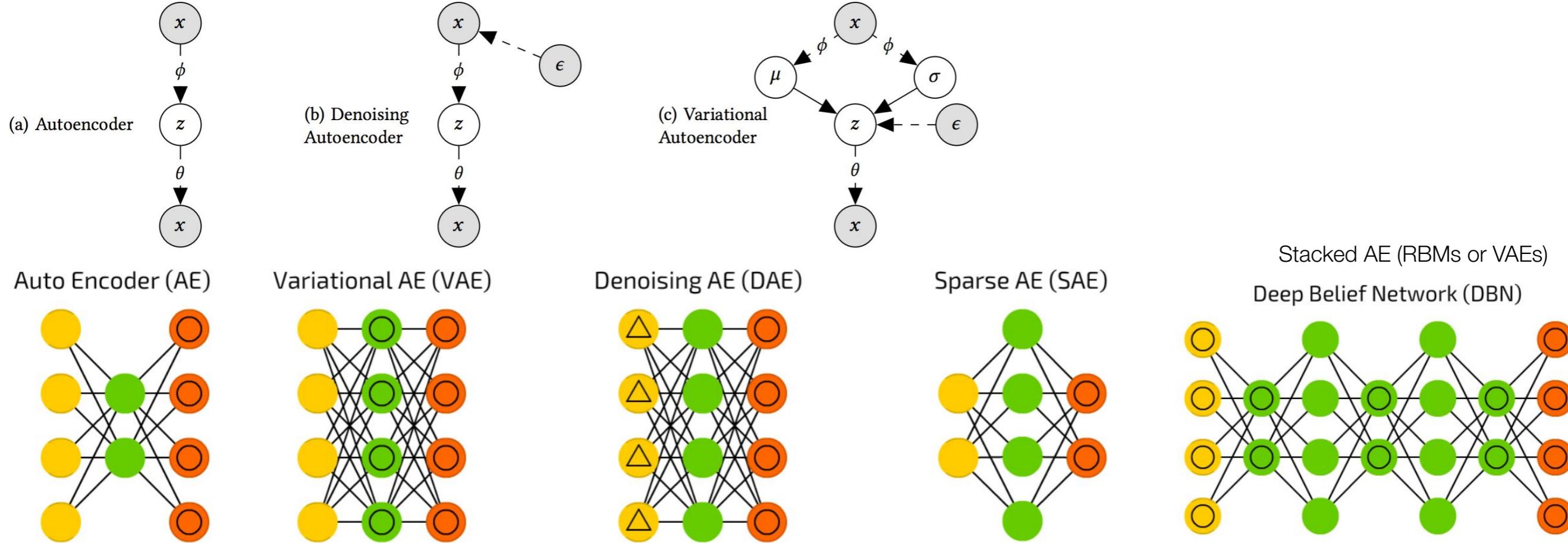
Learning tools

- Denoising (DAE), variational (VAE) and *vanilla* auto-encoders
- Siamese networks
- GANs
- Other objectives



Autoencoders: A taxonomy

[Kramer'91, Salakhutdinov'06, Bengio'09]



- **Pros:** Learn **compact** and **relevant** (undercomplete -bottleneck- and overcomplete [Zazakci'16]) representations minimizing the reconstruction error:
 - DAE inject noise in input
 - VAE in stochastic hidden layer
- **Cons:** Unable to ignore task-irrelevant features (domain randomization or saliency)

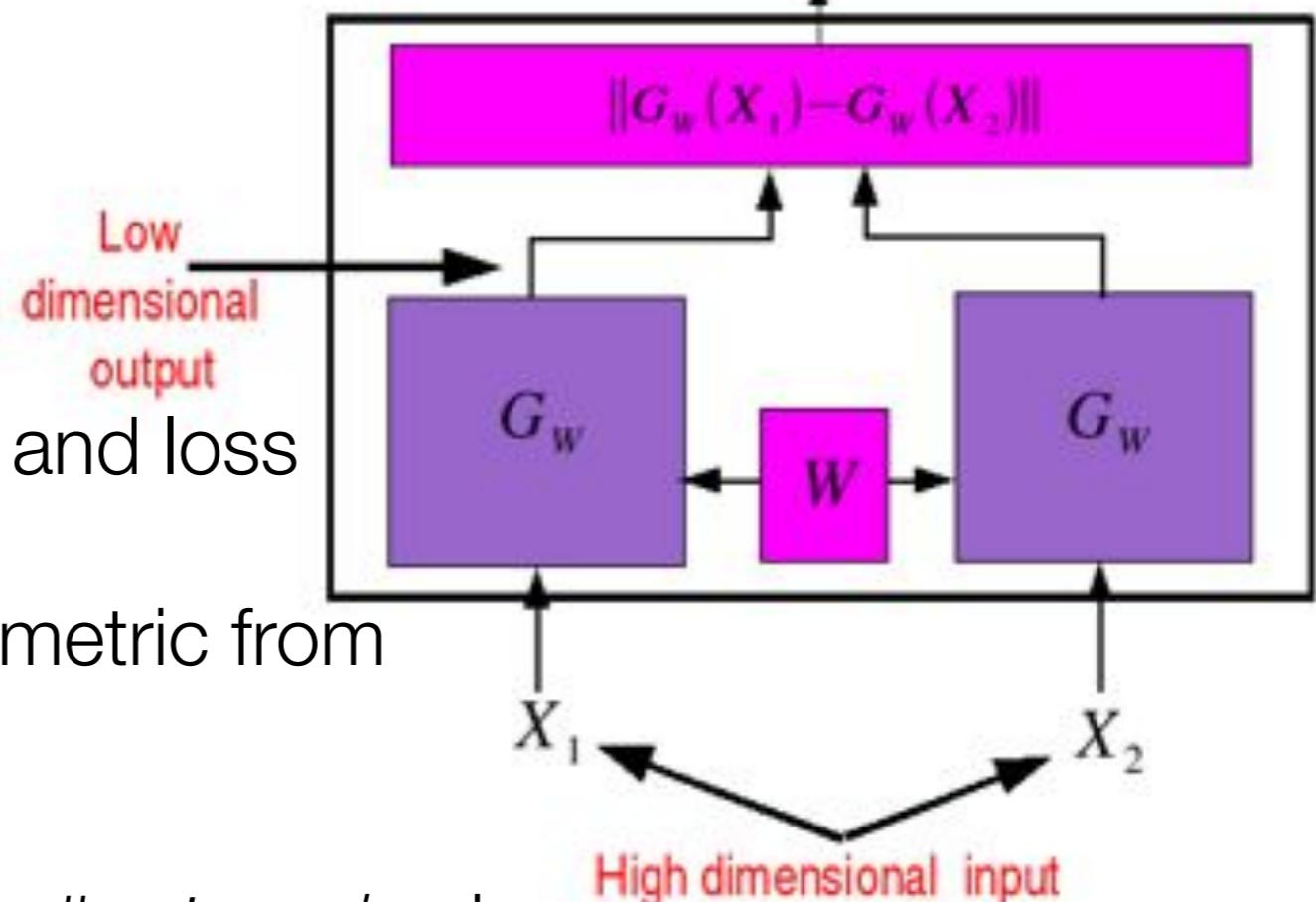
	Input Cell
	Noisy Input Cell
	Hidden Cell
	Probabilistic Hidden Cell
	Spiking Hidden Cell
	Output Cell
	Match Input Output Cell

Siamese Networks

[Chopra'05]

- 2+ identical copies of parameters and loss function
- Discriminatively learns a similarity metric from data
- Suitable:
 - In verification/recognition where #categories is large and unknown
 - Implementing priors (allows multiple passes in parallel)
- Cons: network size doubles

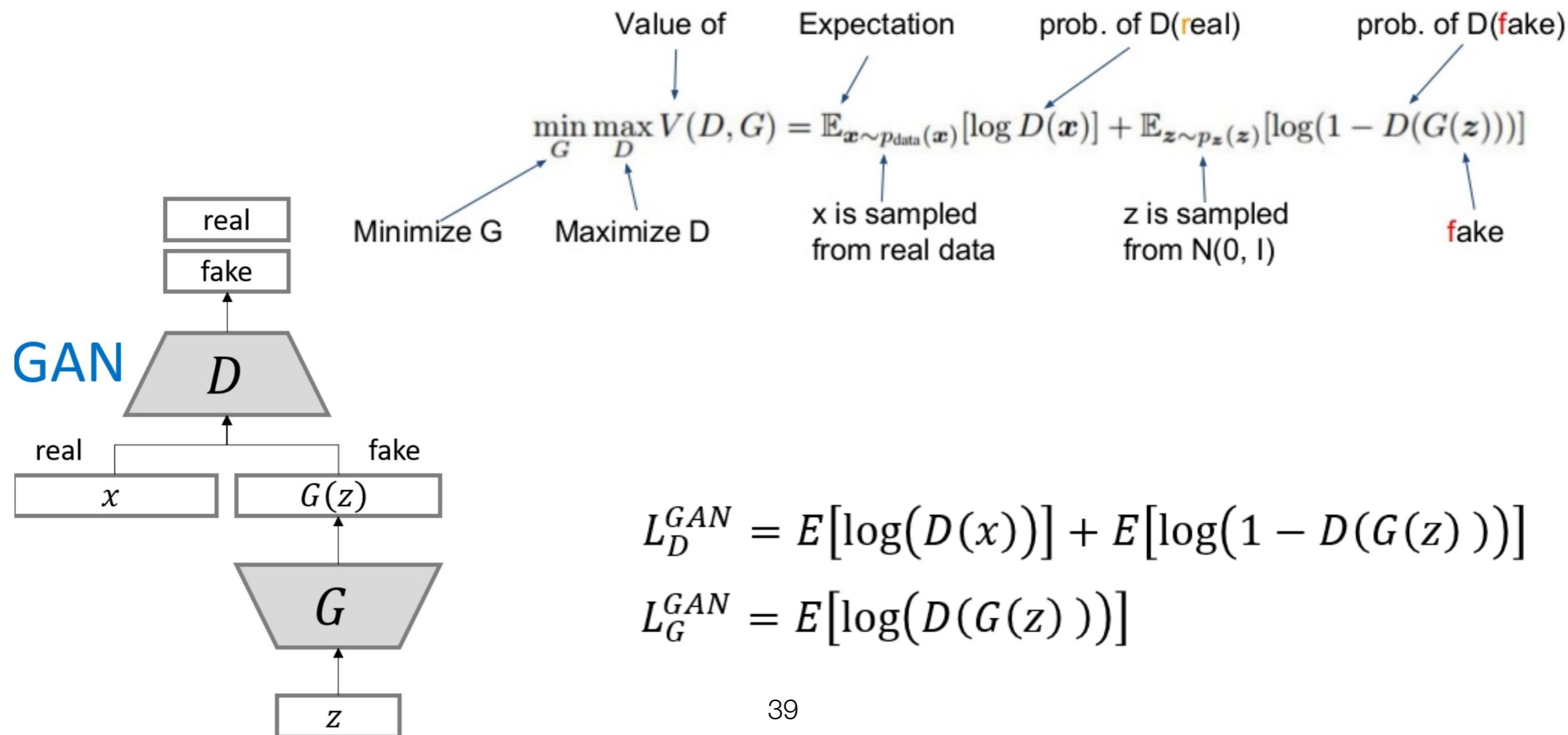
Energy function btw outputs: $E(W, X_1, X_2)$



Generative Adversarial Networks (GANs)

[Goodfellow'14]

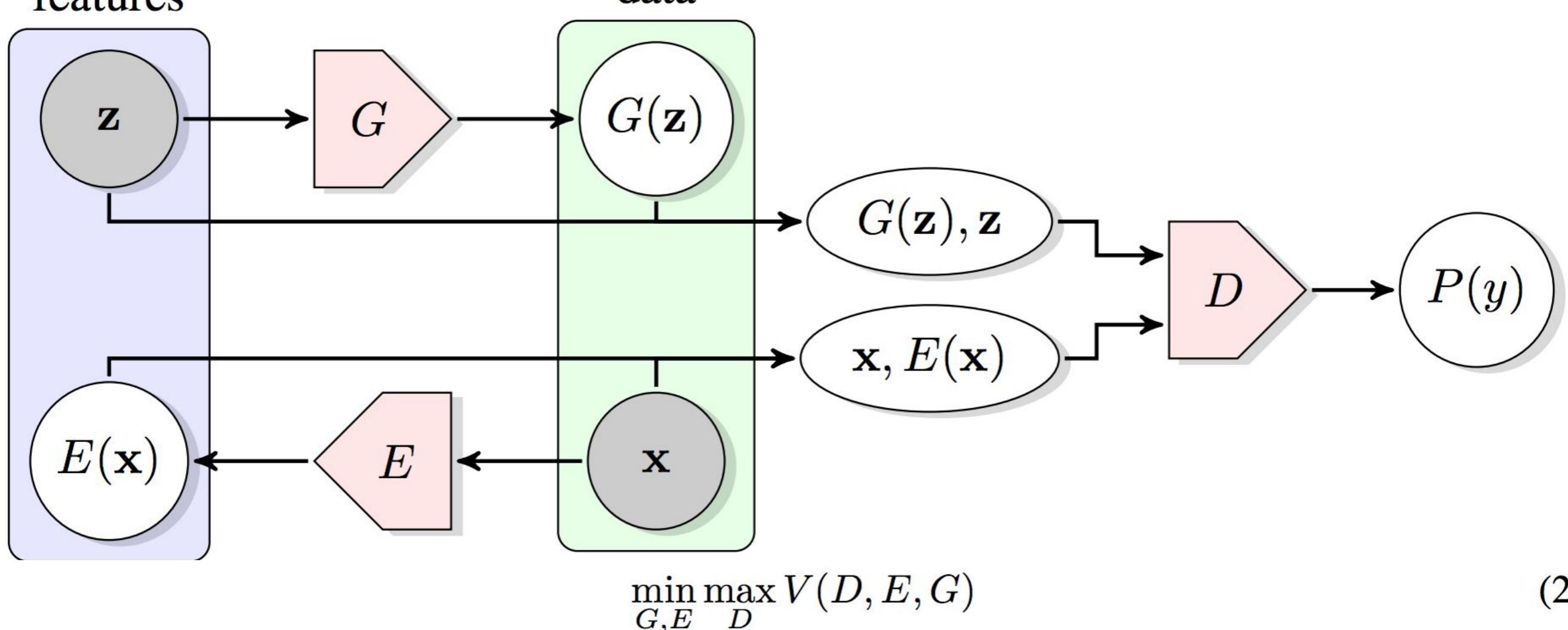
- A two-player adversarial (minimax) game with value function $V(G, D)$ (*1: fake, 0: real input*):



BiGAN: Bidirectional Generative Adversarial Network (= ALI)

features

[Donahue'16, Dumoulin'16]



where

$$V(D, E, G) := \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}}} \left[\underbrace{\mathbb{E}_{\mathbf{z} \sim p_E(\cdot | \mathbf{x})} [\log D(\mathbf{x}, \mathbf{z})]}_{\log D(\mathbf{x}, E(\mathbf{x}))} \right] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{Z}}} \left[\underbrace{\mathbb{E}_{\mathbf{x} \sim p_G(\cdot | \mathbf{z})} [\log (1 - D(\mathbf{x}, \mathbf{z}))]}_{\log(1 - D(G(\mathbf{z}), \mathbf{z}))} \right]. \quad (3)$$

Evaluating state representations: the effect of dimensionality

- Few approaches use a *minimal* state dimension
- Larger dimensions often:
 - Improve performance
 - But limit interpretability

Reference	Observation Dimension	State Dimension	Action Dimension	Environment	Data
Priors [Jonschkowski and Brock, 2015]	16*16*3	2	25 discrete	Slot cars, mobile robot localization	Raw images
PVE [Jonschkow et al., 2017]	Unavailable	5	Discrete	Inverted pendulum, ball in cup, cart-pole	Raw images
E2C [Watter et al., 2015]	40*40*3	8	Discrete	Agent with obstacle	Raw images
E2C [Watter et al., 2015]	48*48*3	8	Discrete	Inverted pendulum	Raw images
E2C [Watter et al., 2015]	80*80*3	8	Discrete	Cart-pole	Raw images
E2C [Watter et al., 2015]	128*128*3	8	Discrete	3 link arm	Raw images
[van Hoof et al., 2016]	20*20*3	3	Continuous	Pendulum swing-up	Raw images
[van Hoof et al., 2016]	228	3	Continuous	Real-robot manipulation task	Tactile data
ML-DDPG [Munk et al., 2016]	18 or 24	6	2 discrete	2 link arm	Joint position
ML-DDPG [Munk et al., 2016]	192 or 308	96	36 discrete	Octopus	Joint position
[Finn et al., 2015]	240*240*3	32	Continuous	Robotics manipulation tasks	Raw images
DVBF [Karl et al., 2016]	16*16*3	3	Unavailable	Pendulum	Raw images
DVBF [Karl et al., 2016]	16*16*3	2	Unavailable	Bouncing ball	Raw images
DVBF [Karl et al., 2016]	16*16*3	12		2 bouncing balls	Raw images
[Goroshin et al., 2015]	3 frames of 32*32	2	2 discrete	NORB dataset	Raw images
ICM [Pathak et al., 2017]	42*42*3	3	4 discrete	3D VizDoom navigation game	Raw images
ICM [Pathak et al., 2017]	42*42*3	2	14 discrete	Mario Bros	Raw images
[Shelhamer et al., 2017]	Unavailable	Un.	Unavailable	Atari	Raw images
VPN [Oh et al., 2017]	3*10*10	Un.	4 discrete	2D navigation	Raw images
VPN [Oh et al., 2017]	4*84*84	Un.	4 discrete	Atari	Raw images
[Curran et al., 2016]	Unavailable	4	5 discrete	Mountain car 3D	Unavailable
[Curran et al., 2016]	Unavailable	12	243 discrete	6 link swimmer	
rwPCA [Parisi, 2017]	21*21*3	Auto.	2 continuous	Picking a coin and putting it on a goal	Raw images
[Parisi, 2017]	20	Auto.	2 continuous	Hit a ball with a ping-pong paddle	Position of objects
[Magrans de Abril and Kanai, 2018]	40	2	Continuous	Explore a 3 room simulated 2D map	Position of agent
[Duan, 2017]	240*240*4	512	4 continuous	Poking cube	Raw images + depth

Evaluating learned representations: Environments

- (Inverted) pendulum, cart-pole
- Atari games (mostly 2D), VizDoom (3D)
- Octopus arms, driving/mountain cars, bouncing ball
- Mobile robots, labyrinths, navigation grids
- **Robotics** manipulation skills:
 - Grasping, balancing a pendulum, poking, pushing a button

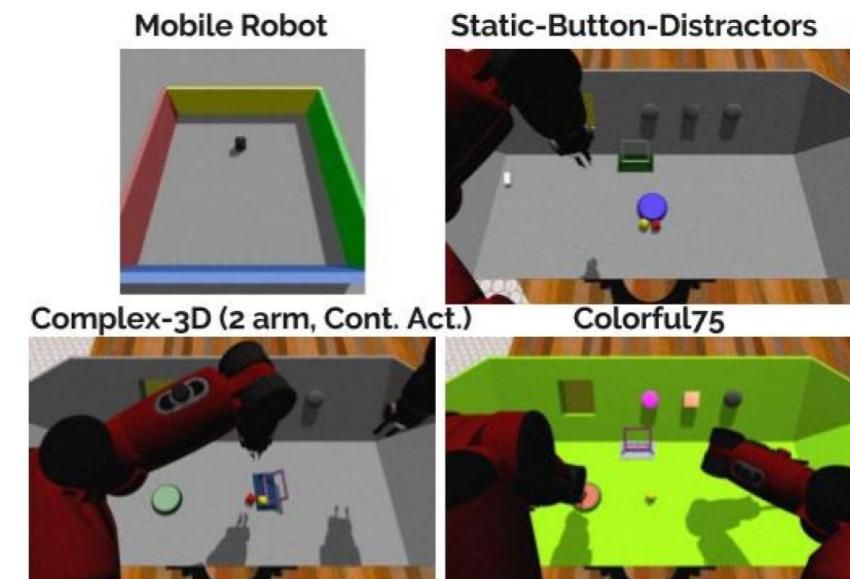
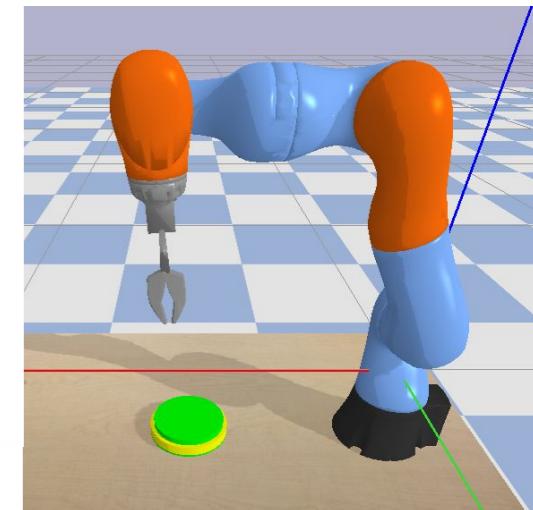


Fig. 3. A sample of each dataset (1-4), created for our benchmark

Figure : Ball-in-cup skill, which the Meka acquired through reinforcement learning



Evaluating state representations: Metrics

Metric	Evaluation target	Application context
Task performance [Jonschkowski and Brock, 2015, Jonschkowski et al., 2017, Munk et al., 2016, van Hoof et al., 2016, Finn et al., 2015, Pathak et al., 2017, Shelhamer et al., 2017, Oh et al., 2017, Parisi et al., 2017, Assael et al., 2015]	Assesses that the information needed to solve a task is contained in s_t	Reinforcement learning
Disentanglement metric score [Higgins et al., 2016]	Measures the disentanglement of the latent factors	Transfer learning, object recognition. Assumes generative factors are known and interpretable
Distortion [Indyk, 2001]	Measures the preservation of local and global geometry coherence	Unsupervised representation learning
NIEQA (Normalization Independent Embedding Quality Assessment) [Zhang et al., 2012]	Measures the local & global neighborhood embedding quality assessment	Manifold learning, not limited to isometric embeddings
KNN-MSE [Lesort et al., 2017]	Measures the degree of preservation of the same neighbors in between the latent space and the ground truth	Unsupervised SLR for robotics

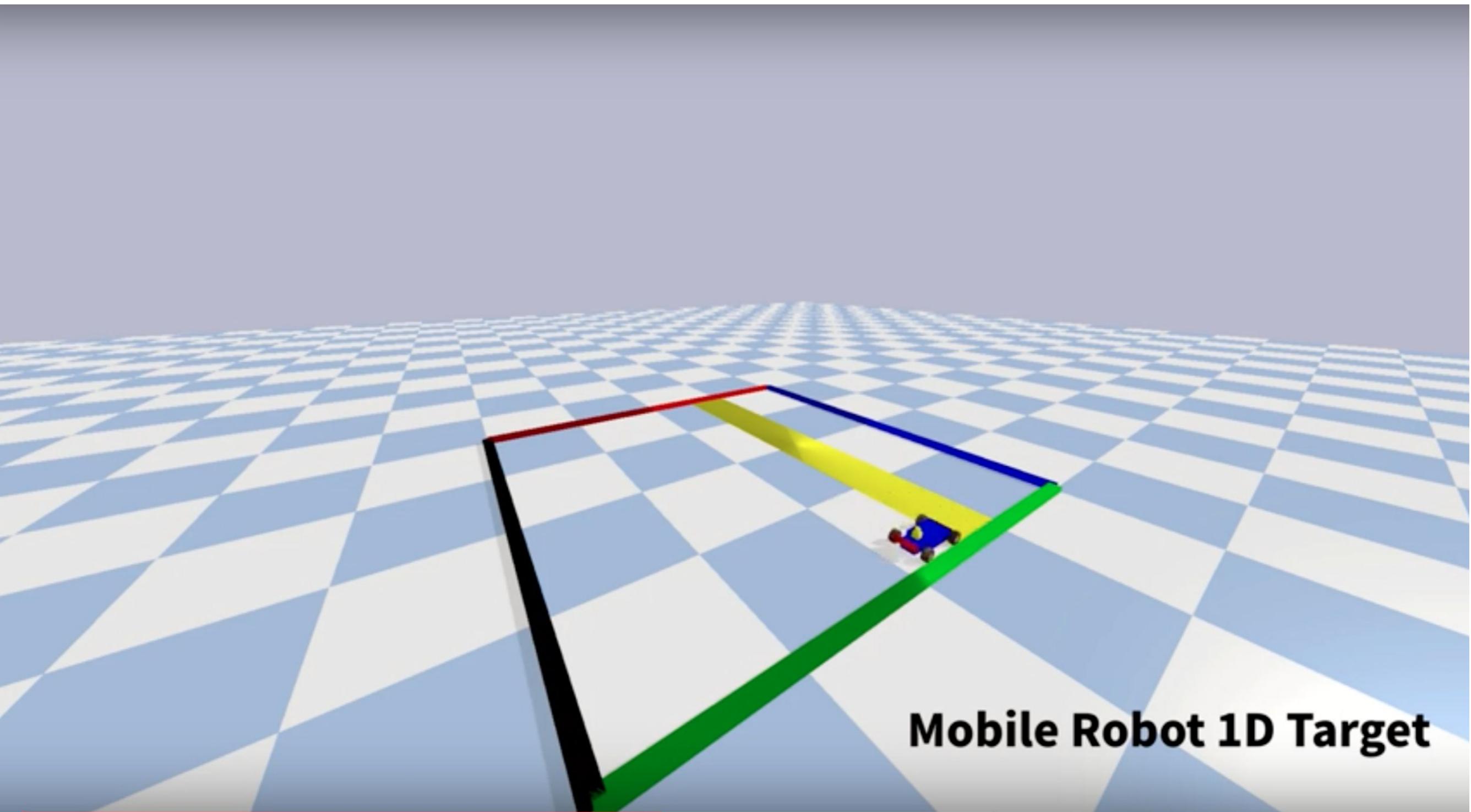
Future challenges/directions

- SRL for autonomous agents
 - Assessment, reproducibility, interpretability (FAT)
 - Automatically decide state dimension
 - Efficient data gathering for SRL by blending in:
 - Learning by demonstration-> with human in the loop!
 - Few-shot learning
 - GANs
 - Intrinsic motivations via Goal Exploration Processes
- => Searching for PhDs on co-supervision!

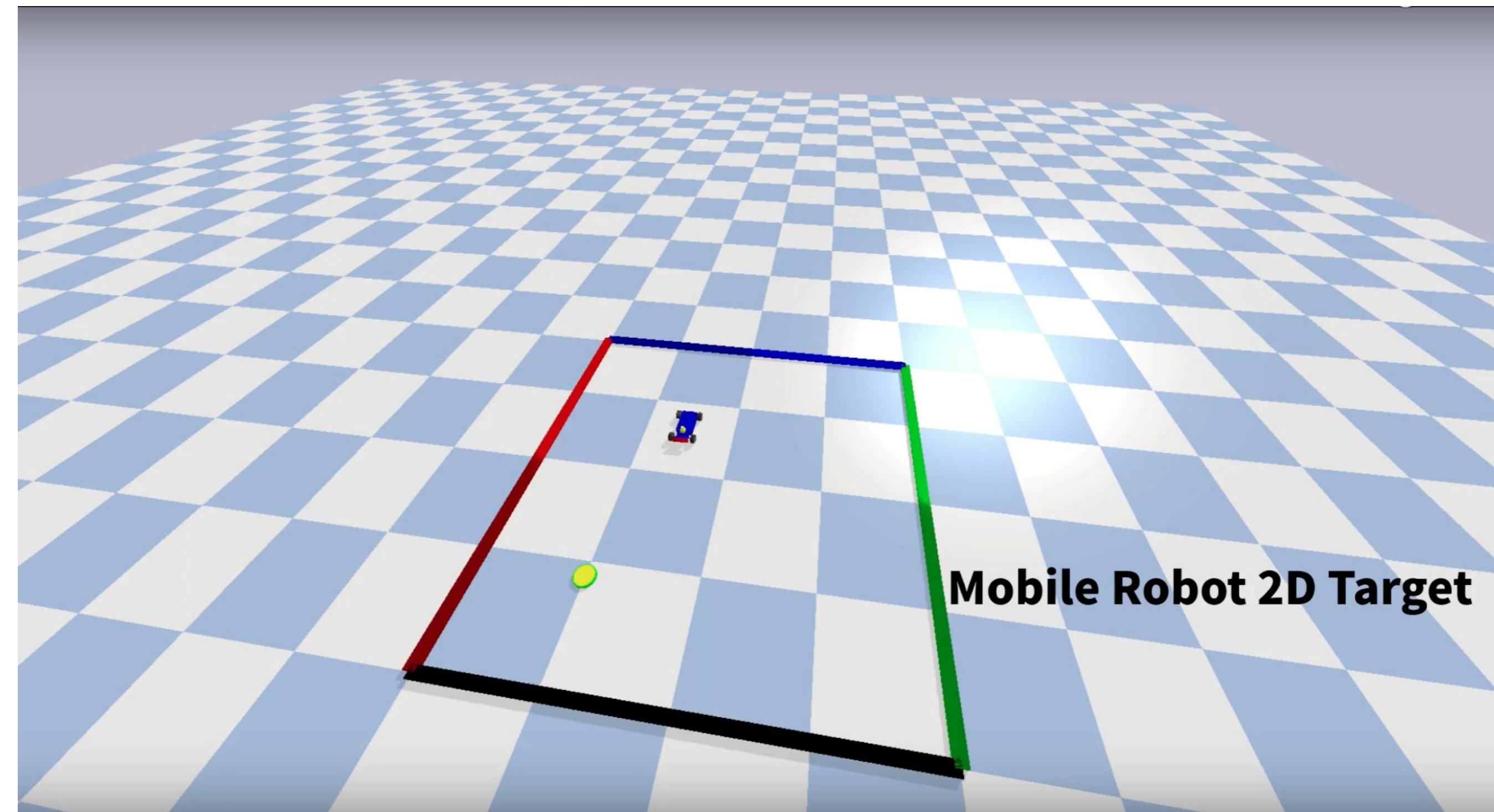
(MAIN) REFERENCES

- [GAN Zoo](#) and [GANs presentation](#)
- [Deep learning Zoo](#)
- Lesort et. al. [Unsupervised state representation learning with robotic priors: a robustness benchmark](#), 2017, [DEMO video](#)
- Lesort et al. [State Representation learning for control: an Overview](#) Neural Networks, Elsevier, 2018.
- Doncieux et al. [Representational Redescription in Open-ended Learning: a Conceptual Framework](#)

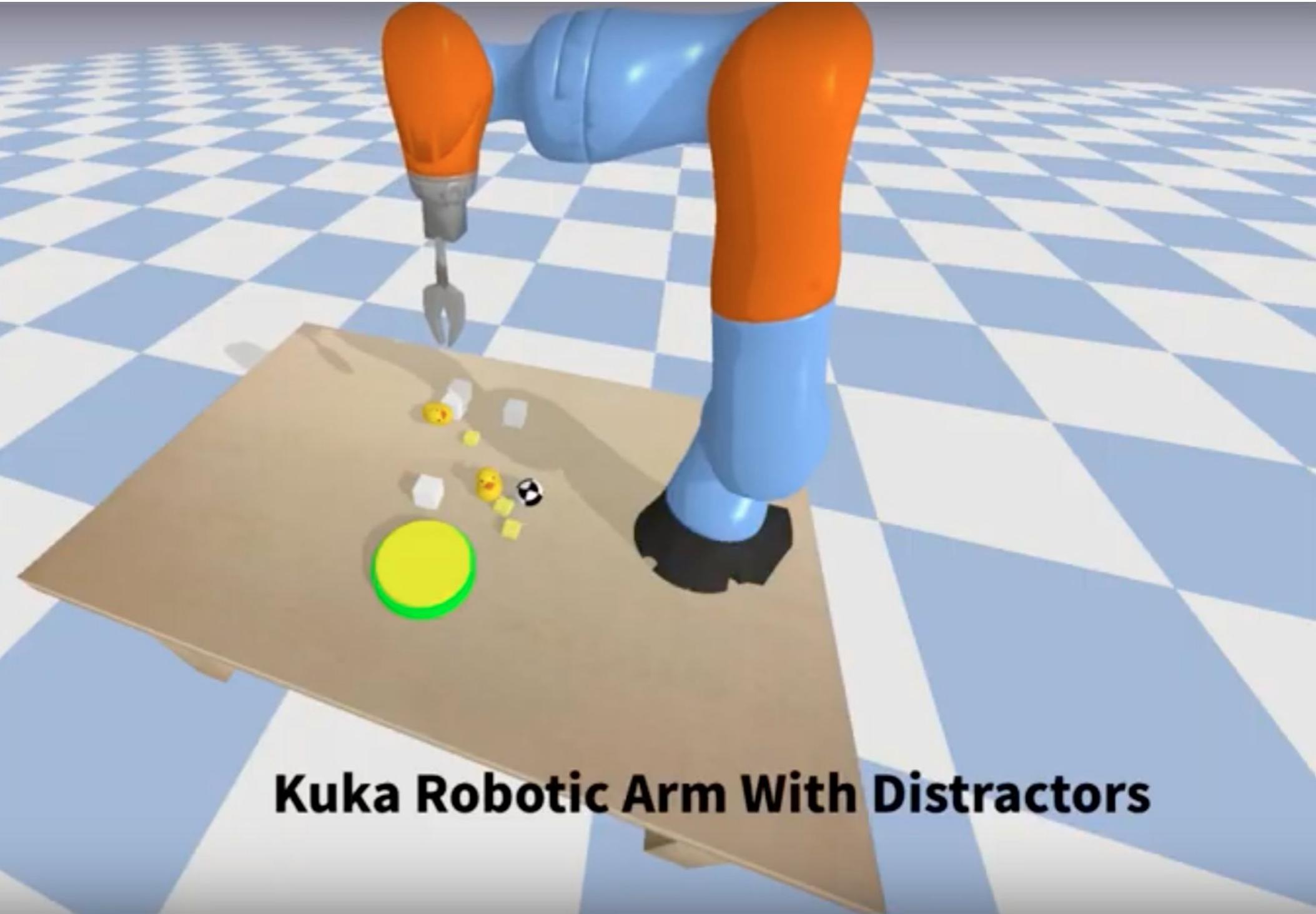
S-RL Toolbox



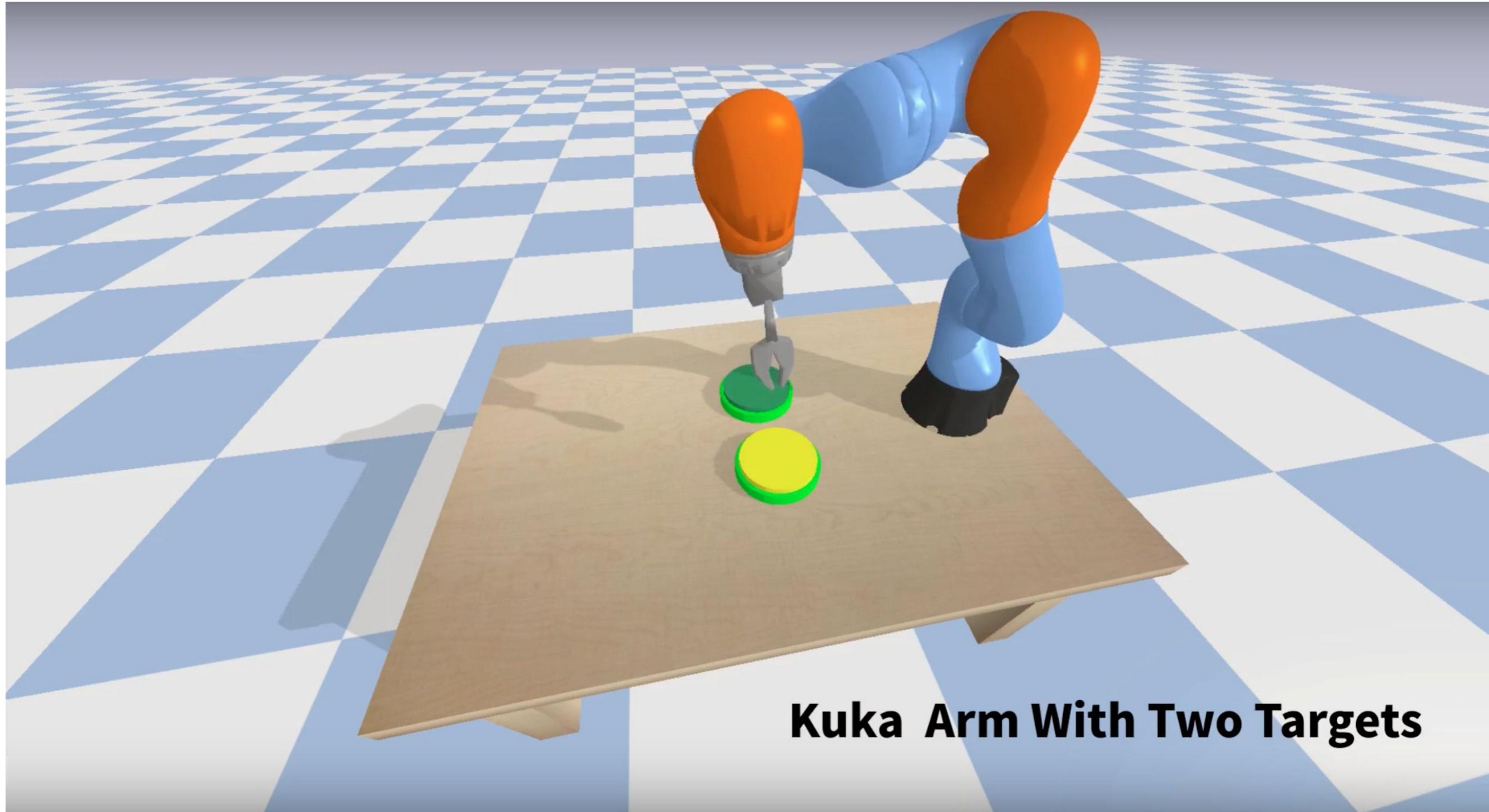
S-RL Toolbox



S-RL Toolbox

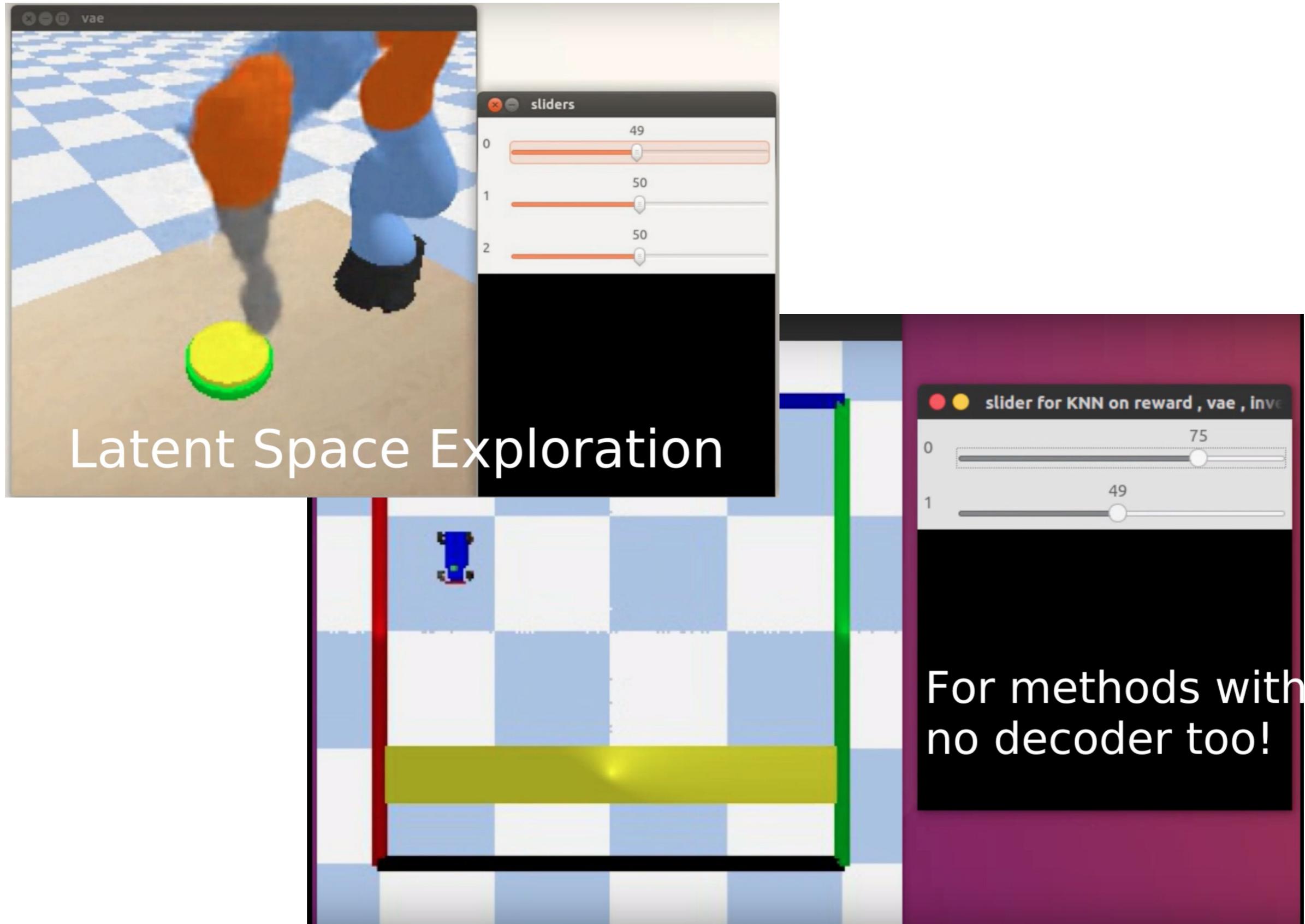


S-RL Toolbox



Kuka Arm With Two Targets

S-RL Toolbox

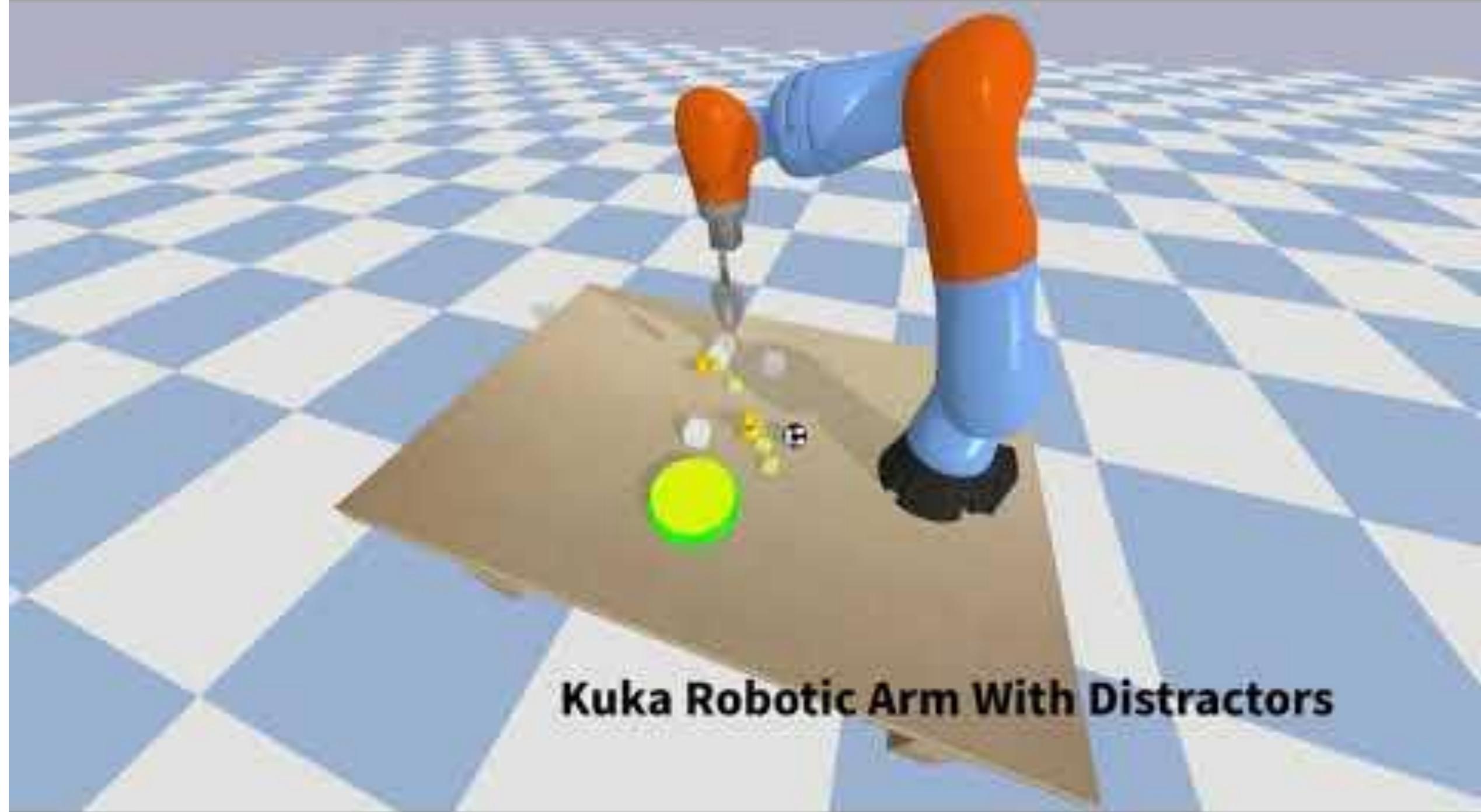


For methods with
no decoder too!

S-RL Toolbox



Real Baxter Robot Static Target



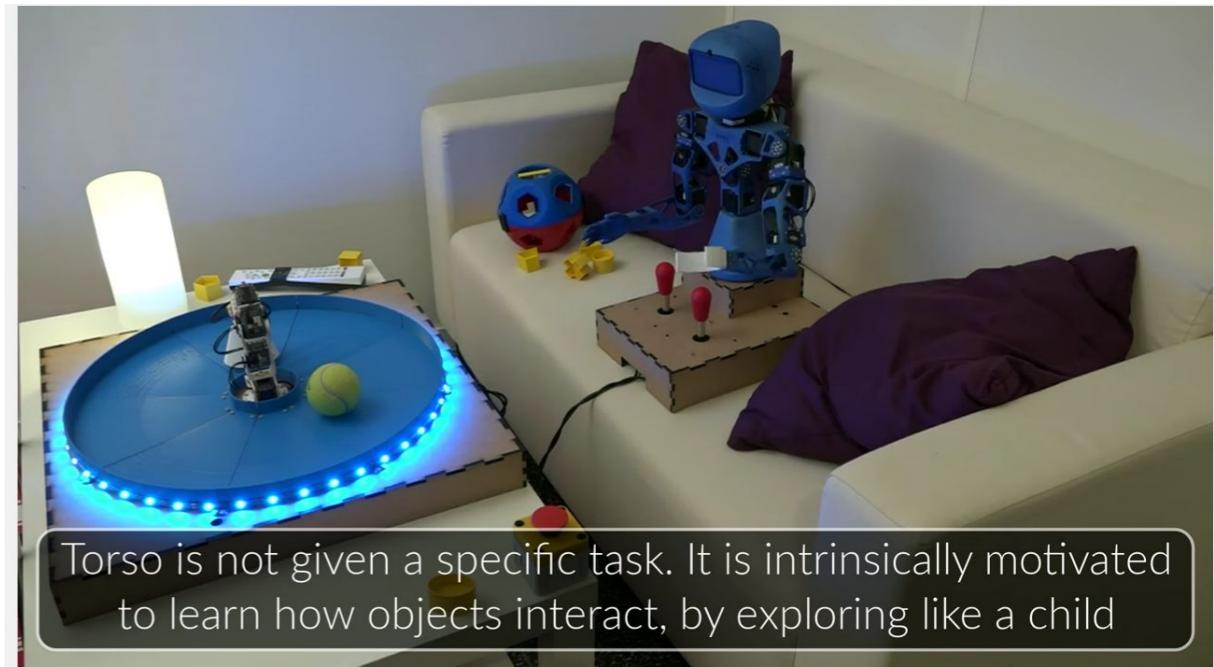
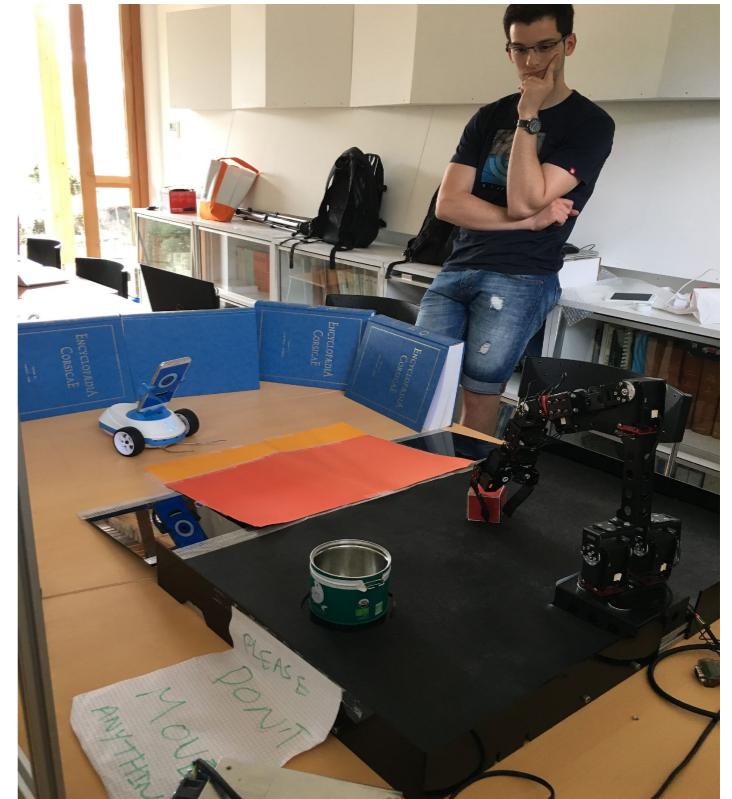
KUKA Robotic Arm With Distractors

Future Perspectives

- More & More complex tasks: Rewards?



Robobo



Torso is not given a specific task. It is intrinsically motivated to learn how objects interact, by exploring like a child

Intrinsically Motivated Multi-Task Reinforcement Learning

Thank you! Project Ideas?

Connect: nataliadiaz.github.io
natalia.diaz@ensta-paris.fr

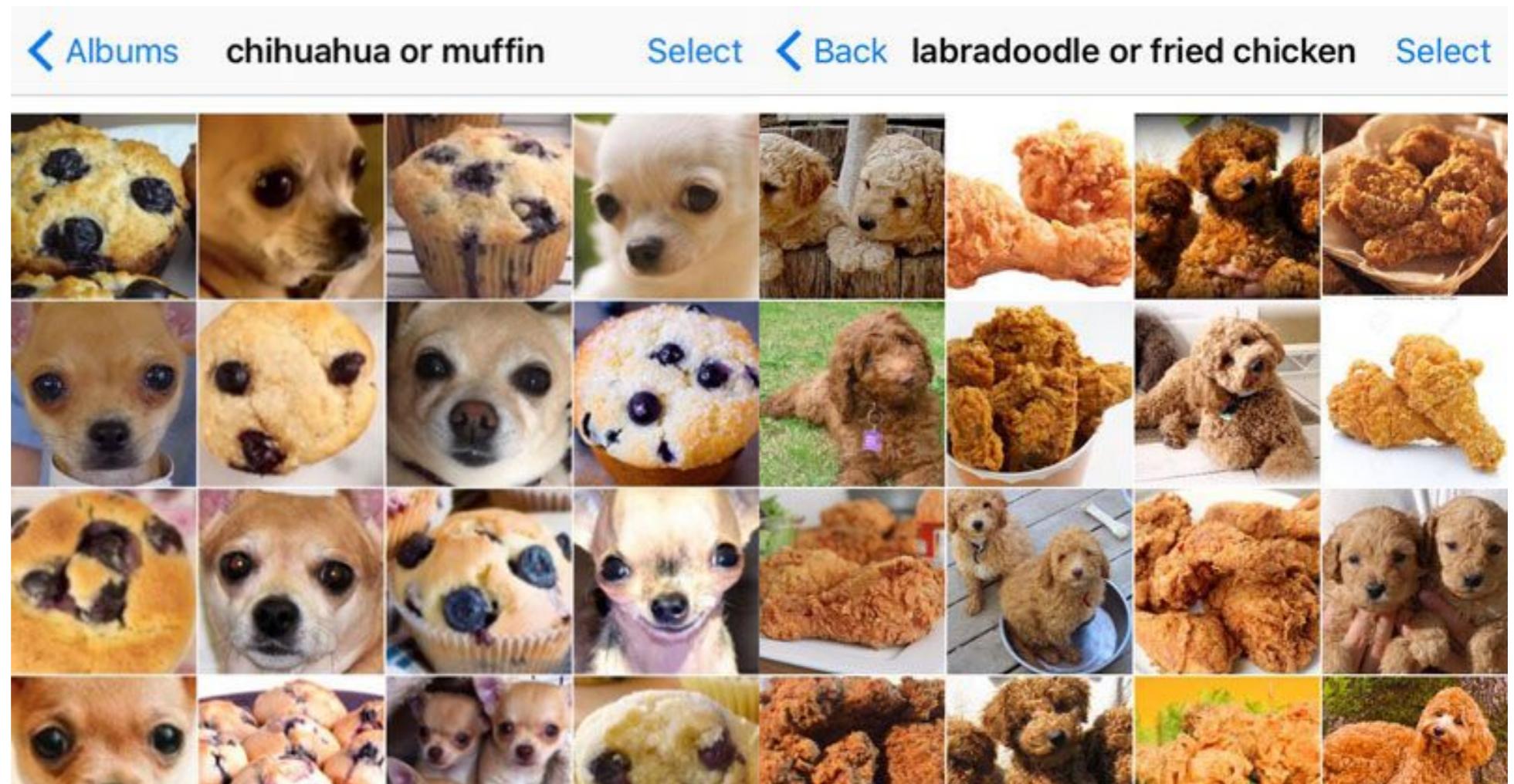


◀ Albums chihuahua or muffin Select Back labradoodle or fried chicken Select

A 4x4 grid of 16 small images. The first column contains three images of muffins with blueberries. The second column contains three images of Chihuahuas. The third column contains three images of muffins with blueberries. The fourth column contains three images of Chihuahuas. The fifth column contains three images of Labradoodles. The sixth column contains three images of fried chicken. The seventh column contains three images of Labradoodles. The eighth column contains three images of fried chicken.

Appendix

Future directions *(or deep learning beyond classifying chihuahuas and blueberry muffins)*



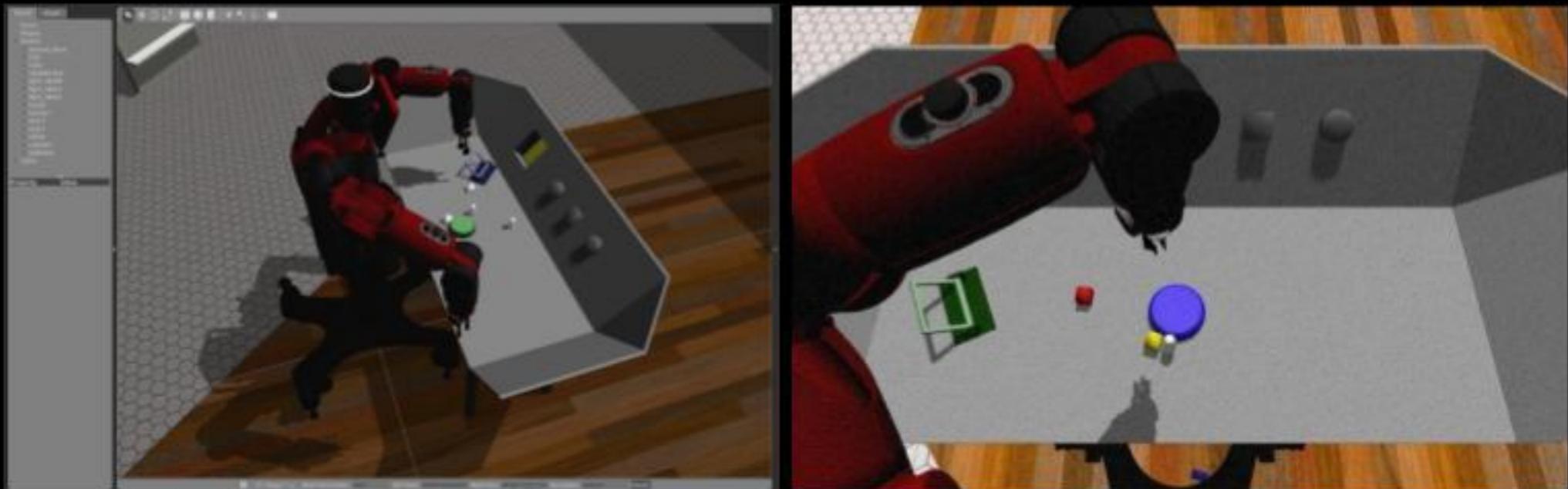
Appendix

*Unsupervised state
representation learning with
robotic priors: a robustness
analysis*

<https://arxiv.org/abs/1709.05185>

State Representation Learning with robotics priors

*Unsupervised state representation learning
with robotic priors: a robustness benchmark*



T. Lesort, M. Seurin, X. Li, N. Díaz-Rodríguez, D. Filliat
U2IS - ENSTA ParisTech / ARMINES, France
www.robotsatdream.eu ICRA 2018

State Representation Learning with robotics priors



The goal is to discover a state space allowing to push a button



State Representation Learning with robotics priors

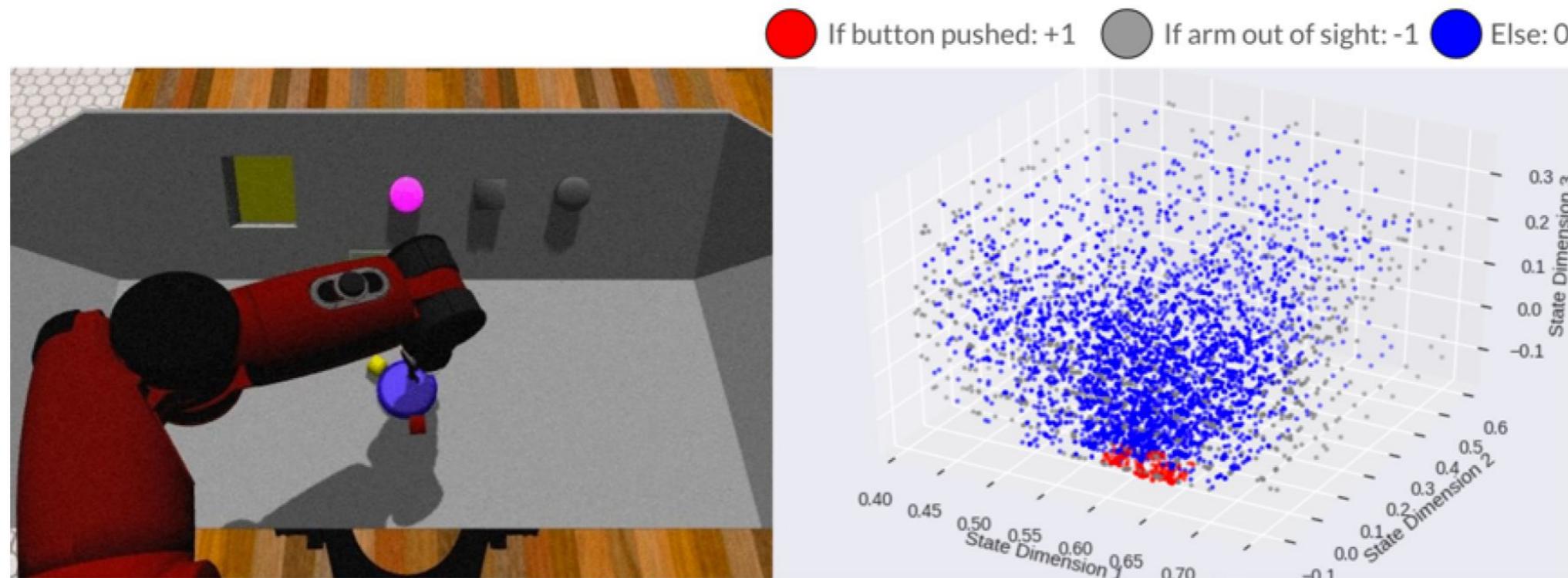
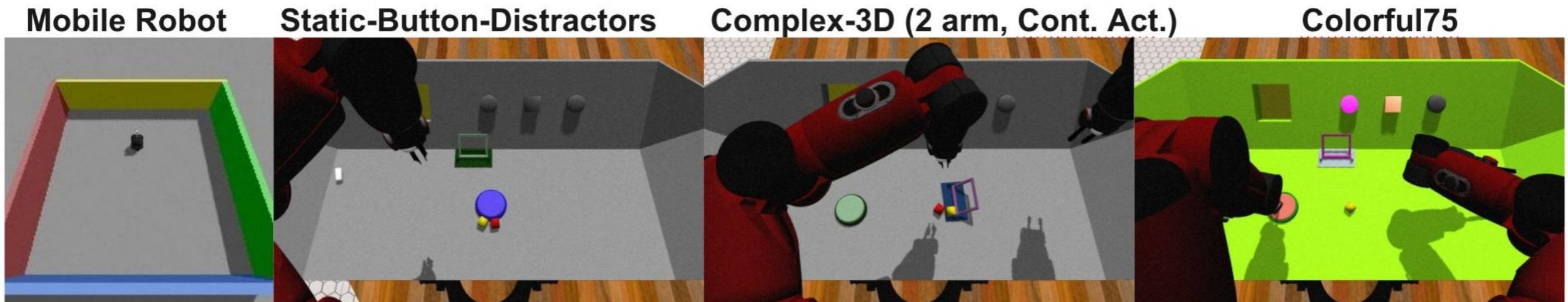


Fig. 2. Left: Baxter's camera view for Static-Button-Distractors dataset 2. Right: Baxter's left hand position ground truth position and its coded reward



State Representation Learning with robotics priors



Figure 4: Nearest neighbor evaluation of robotic priors against supervised learning and Autoencoders on (the domain randomization) Colorful75 dataset using the fixed ref. point prior. The auto-encoder-learnt nearest state's observation is far from the ground truth position.

Criterion	GT	Superv	4 Priors	5 Priors	AE	5 Priors Of
KNN-MSE	0.024	0.03	0.079	0.053	0.099	0.047
NIEQA local	0	0.239	0.66	0.50	0.599	0.52
NIEQA global	0	0.048	0.41	0.20	0.465	0.21

Table 1: *Static-Button-Distractor 3D* (dataset 2) Results. xf means x ResNet frozen layers, GT: (hand position) Ground Truth

T. Lesort et. al. Unsupervised state representation learning with robotic priors: a robustness benchmark, 2017, <https://arxiv.org/abs/1709.05185> DEMO: <https://www.youtube.com/watch?v=wFd0yJuJIQ4>

State Representation Learning with robotics priors



while autoencoders fail completely

Supervised (Robot Hand Position)
record_0/3/frame00112
[0.635 0.453 0.103] d=0.0189
[0.7135355 0.4438927 0.0785464]



Robotic Priors
record_0/20/frame00196
[-0.293 -25.6 -1.181] d=0.0534
[0.74266221 0.34607499 0.05758329]



Denoising Auto-Encoder
record_0/21/frame00213
[0.000 13.4438561662843000 0.000] d=12860275319771226.0000
[0.4410974 0.2269147 -0.05124815]



T. Lesort et. al. Unsupervised state representation learning with robotic priors: a robustness benchmark, 2017, <https://arxiv.org/abs/1709.05185> DEMO: <https://www.youtube.com/watch?v=wFd0yJuJIQ4>

State Representation Learning with robotics priors

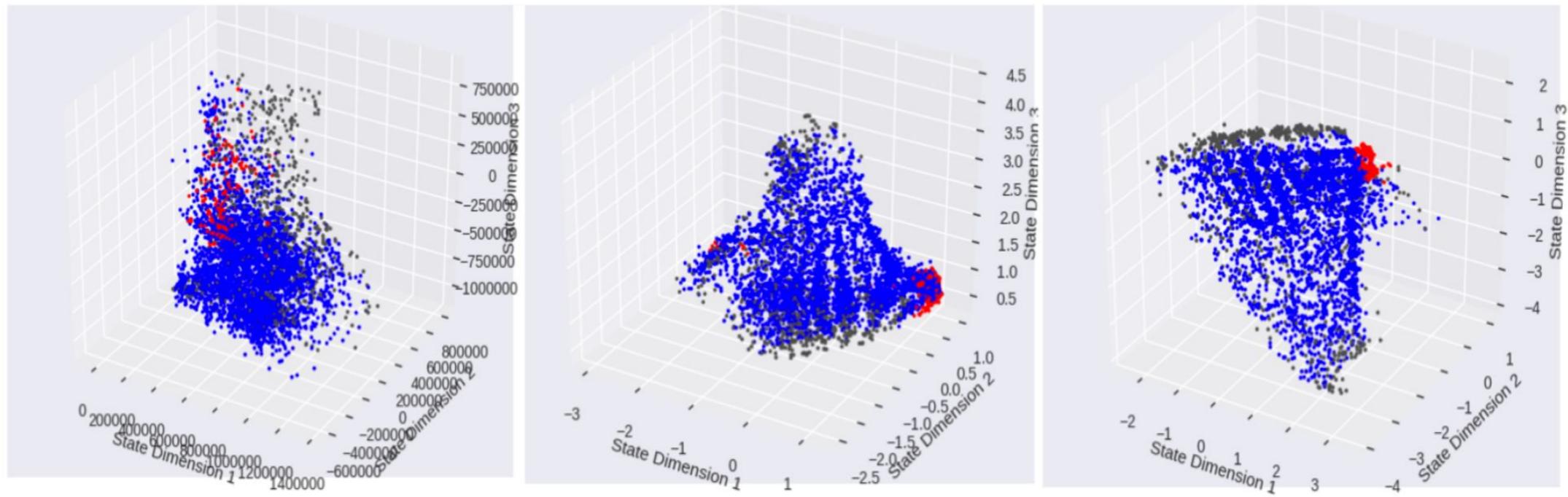


Fig. 4. Learned state space on Static-Button-Distractors (dataset 2): Left: Denoising Autoencoder. Middle: 4 Priors. Right: 5 Priors. A red reward (value +1) state means the button is being pushed, gray (-1) if the hand is out of sight, and blue (reward 0) if hand is elsewhere.

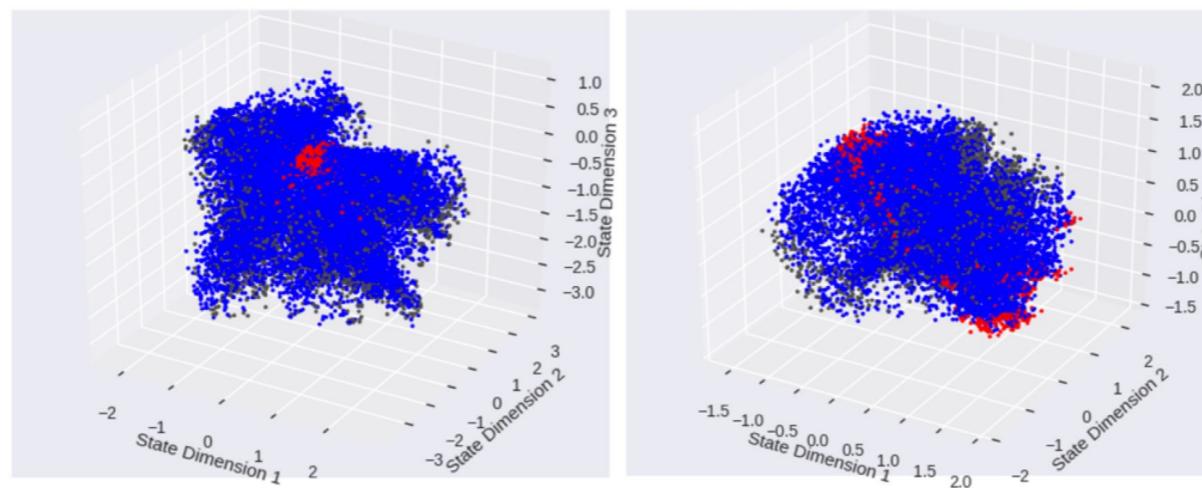


Fig. 6. Results of the 5 priors on *colorful75* (dataset 4). Left: 5 priors using button position as ref. point. Right: 5 priors using starting hand position as ref. point.

Robotic Priors

[Jonschkowski et. al. 2015]

Effect of distractors and domain randomization without Ref. point prior

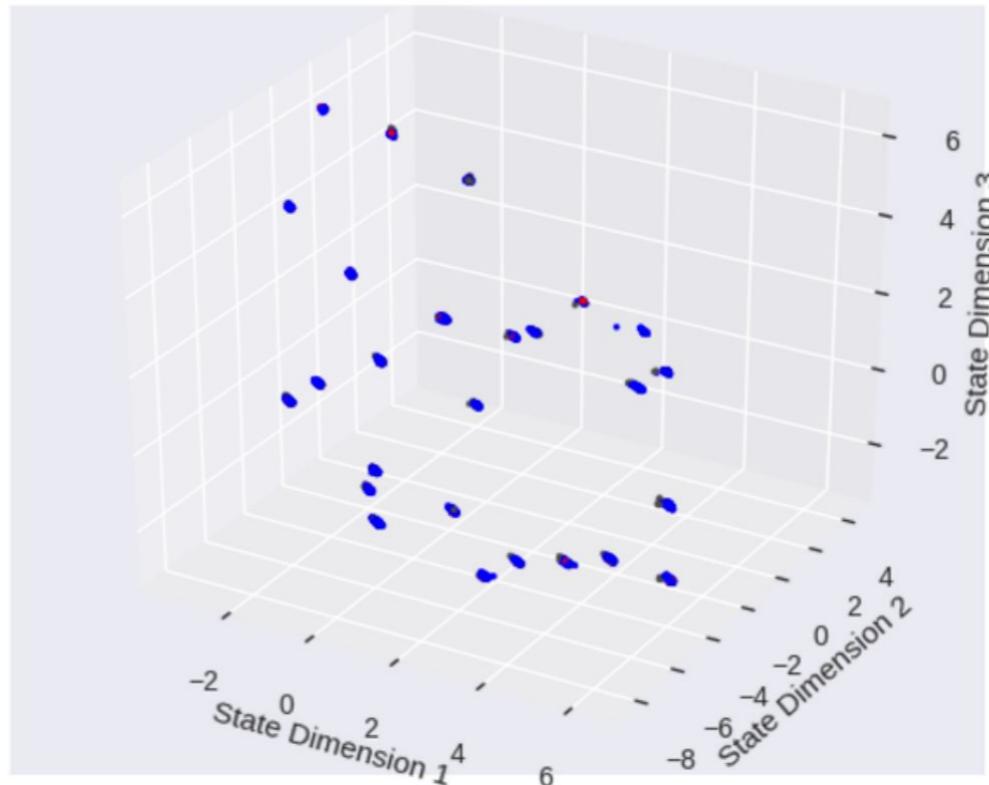
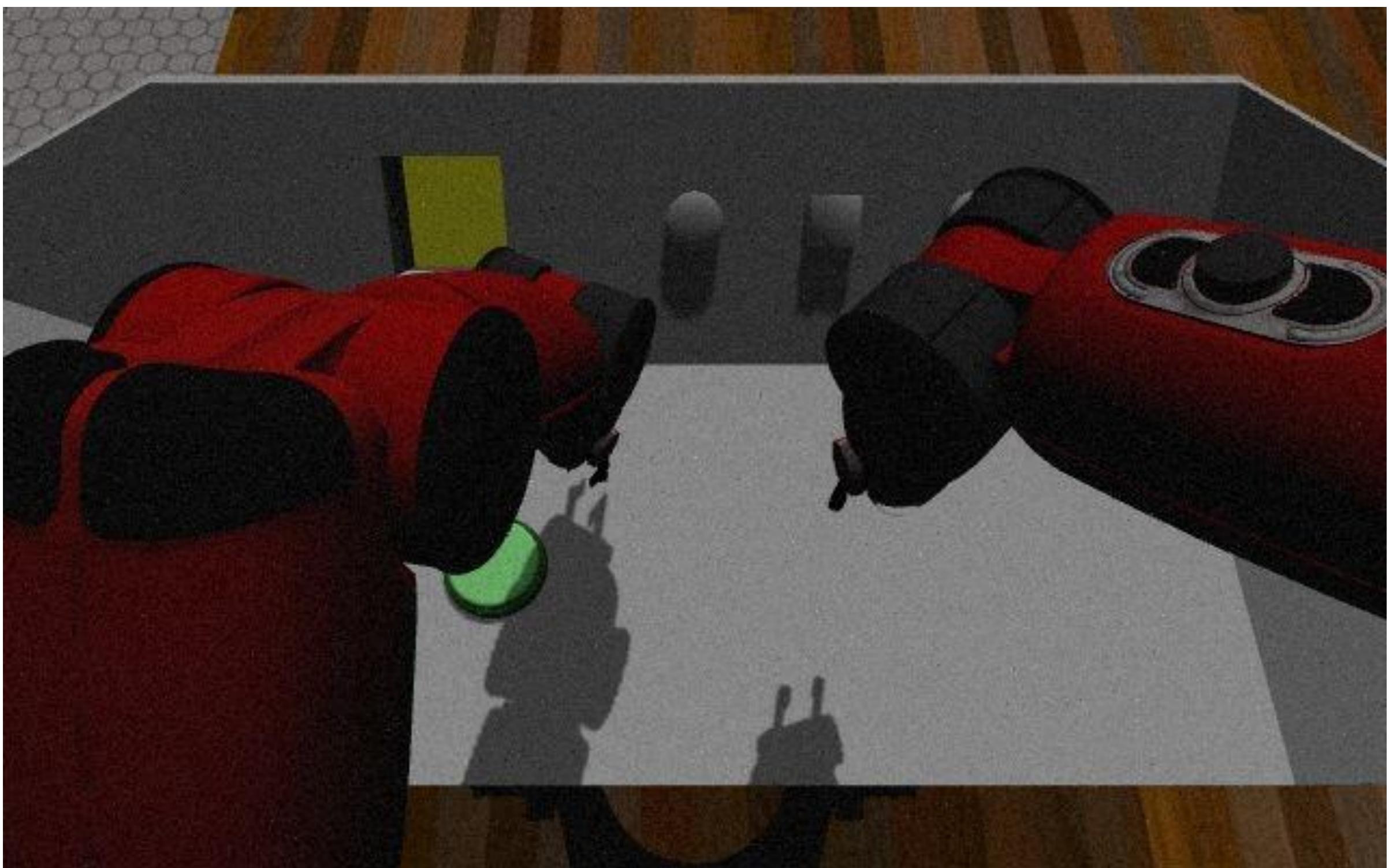


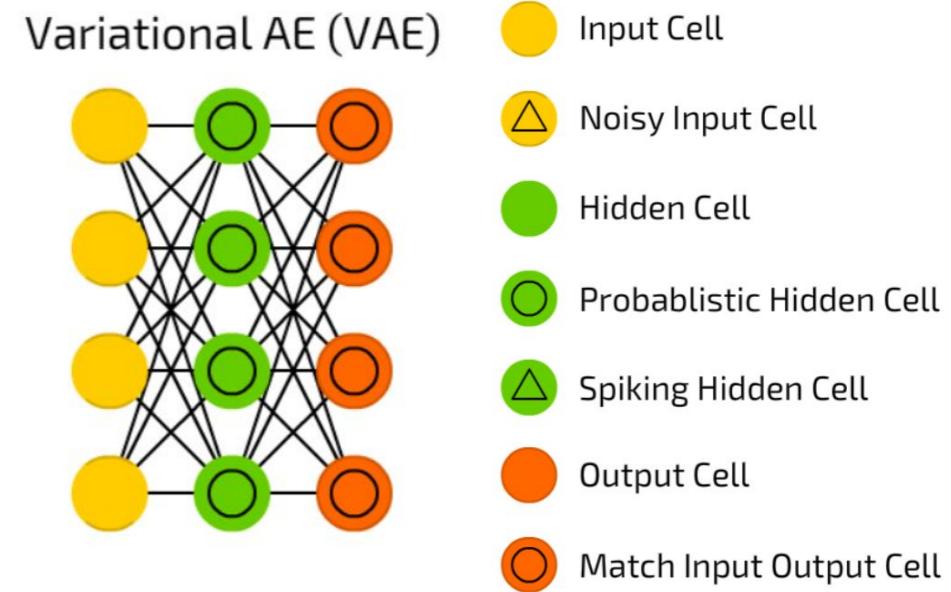
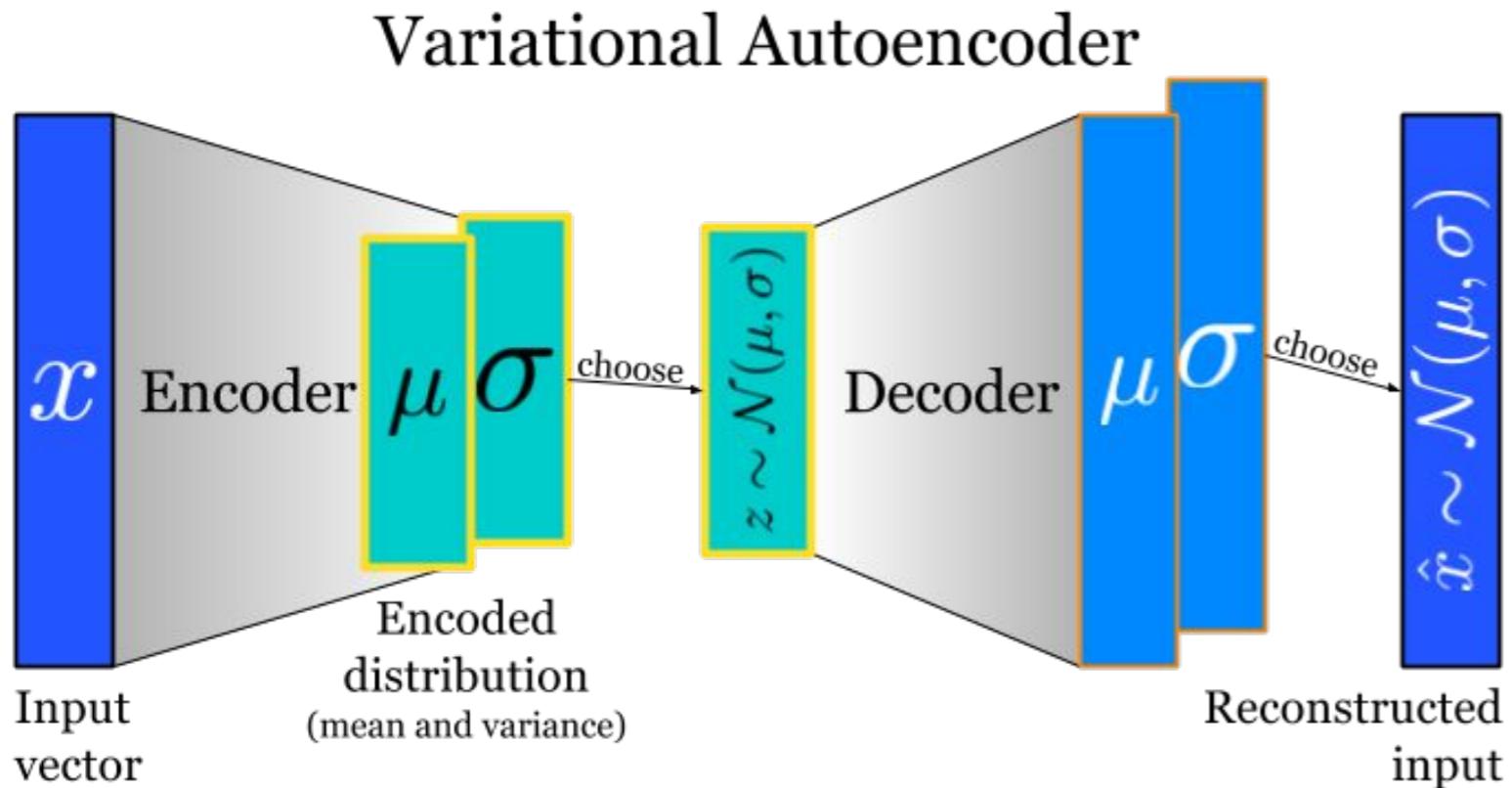
Fig. 5. Effect of static distractors in dataset 3 on the 4 priors approach learned state space.



CONTRIBUTIONS:

- *Robotic Priors ignore distractors and learn representations more relevant to the task than autoencoders*
- *Performance is close to the one learnt with supervision*
- *A quantitative and qualitative evaluation metric based on KNN-MSE*
- *Vulnerabilities of 4 original robotic priors are overcome with a contributed 5th fixed reference point prior*

Generative models: Variational autoencoders



The objective of VAE is to maximize the following variational lower bound with respect to the parameters θ and ϕ .

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \quad (2)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \quad (3)$$

Inverse model example: Intrinsic Curiosity Module (ICM)

Curiosity-driven Exploration by
Self-supervised Prediction

Deepak Pathak Pulkit Agrawal Alexei A. Efros Trevor Darrell

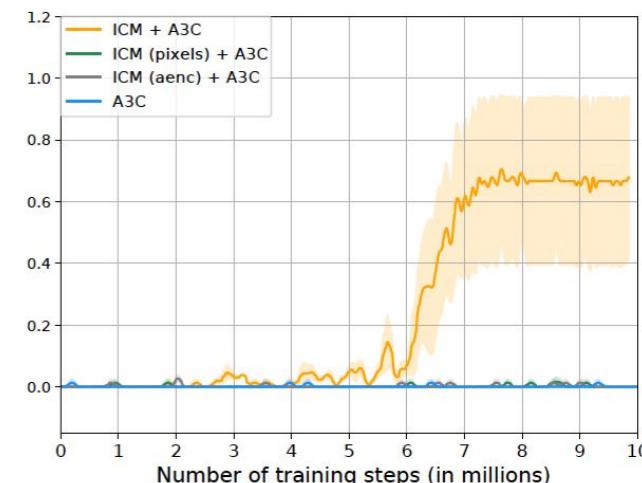
University of California, Berkeley

ICML 2017

[Download Paper]

[Github Code]

- *Curiosity*: error in an agent's prediction of the consequence of its own actions in a visual feature space (learned by a self-supervised inv. dynamics model).
 - Allows reward to be optional
 - Bypasses difficulties of predicting pixels
- Cons: Not tested on robotics 3D images



(c) “very sparse reward” setting

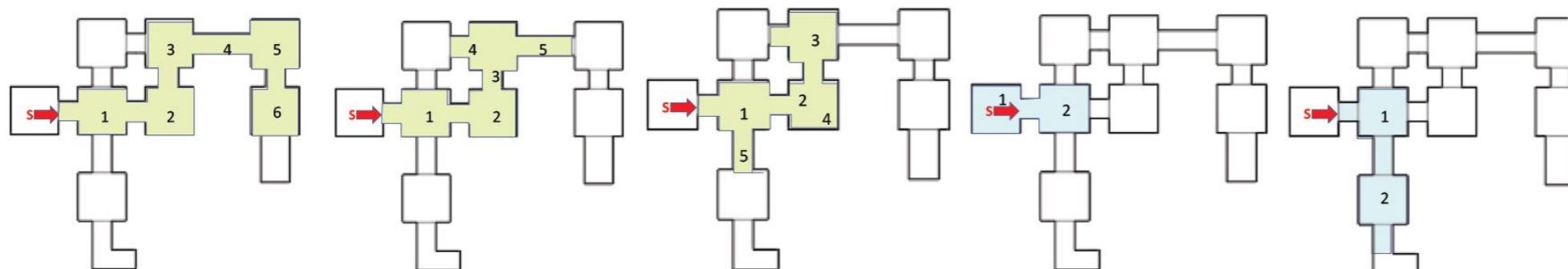
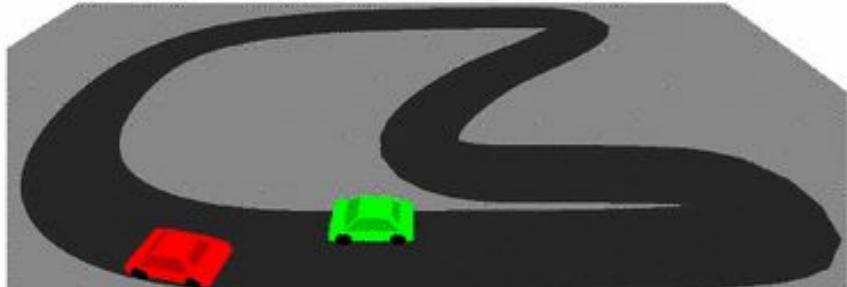


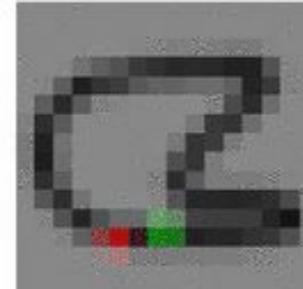
Figure 7. Each column in the figure shows the coverage of an agent by coloring the rooms it visits during 2100 steps of exploration. The red arrow shows the initial location and orientation of the agent at the start of the episode. The first three (in green) and the last two columns (in blue) show visitation of curious (ICM) and randomly exploring agents respectively. The results clearly show that the curious agent trained with intrinsic rewards explores a significantly larger number of rooms as compared to a randomly exploring agent.

Robotic Priors

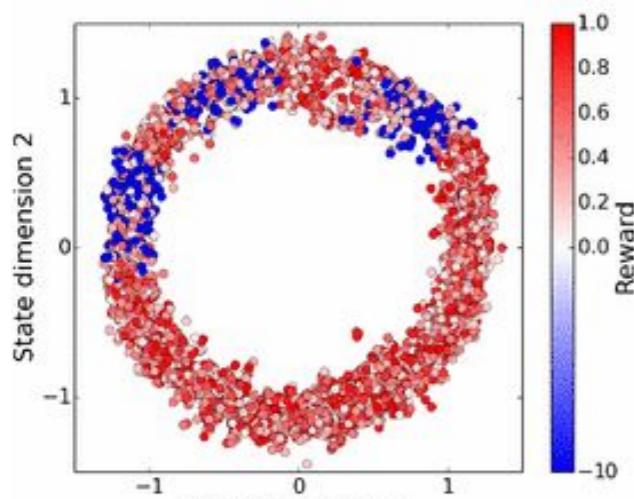
[Jonschkowski et. al. 2015]



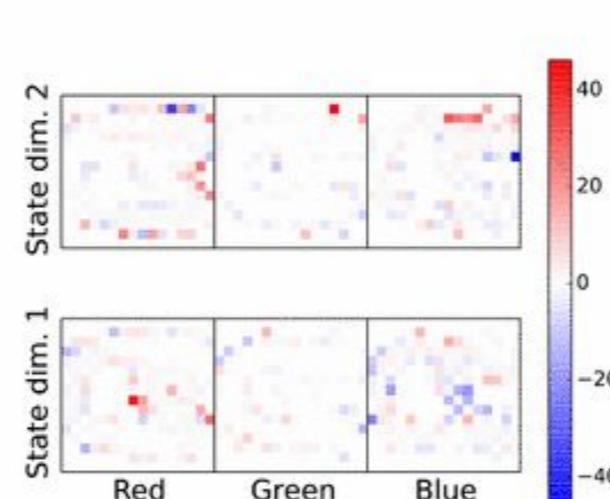
(a)



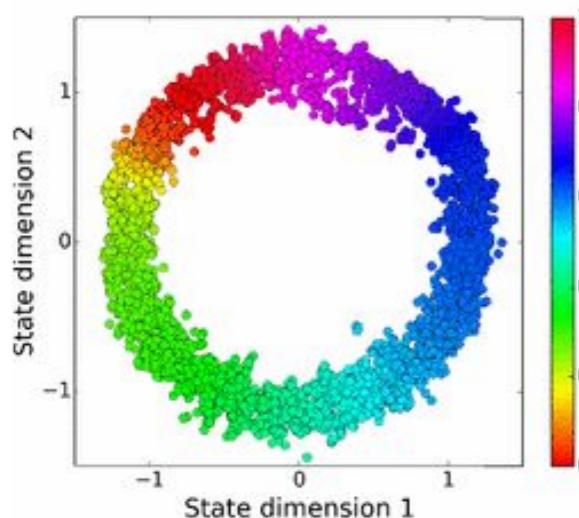
(b)



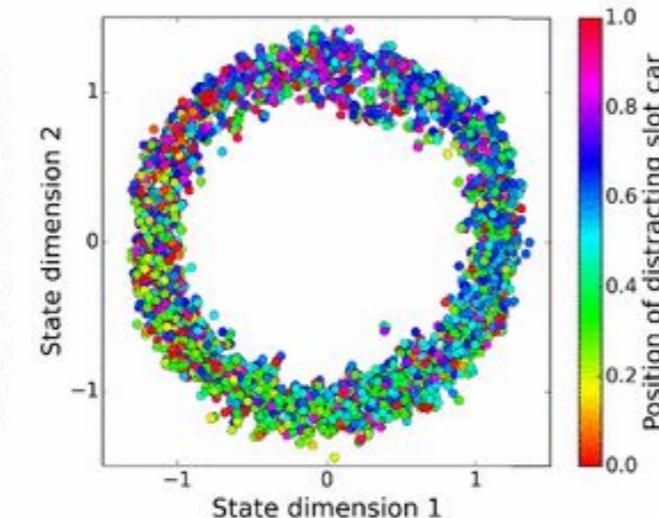
(c)



(d)



(e)



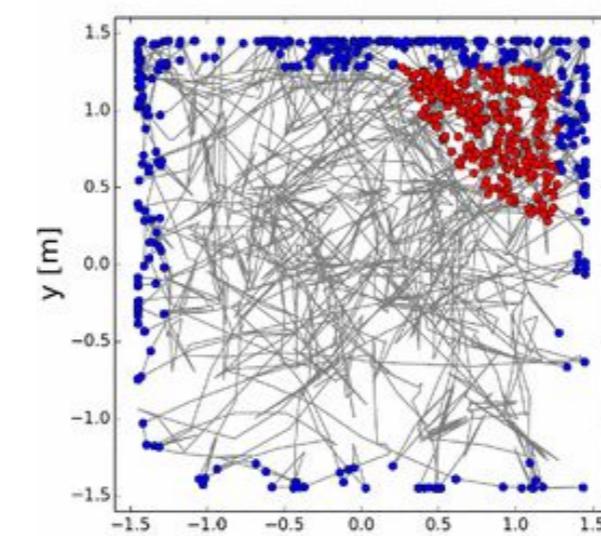
(f)



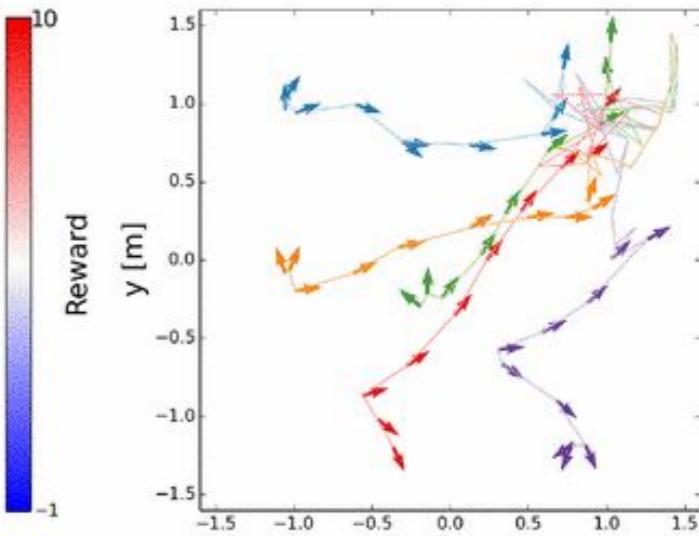
(a)

(b)

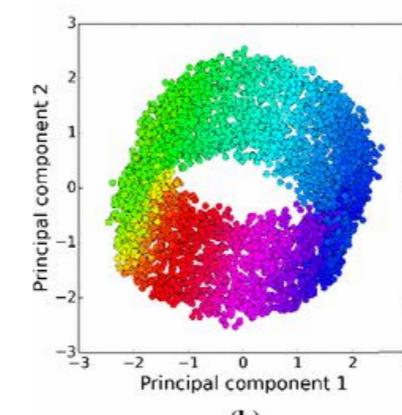
(c)



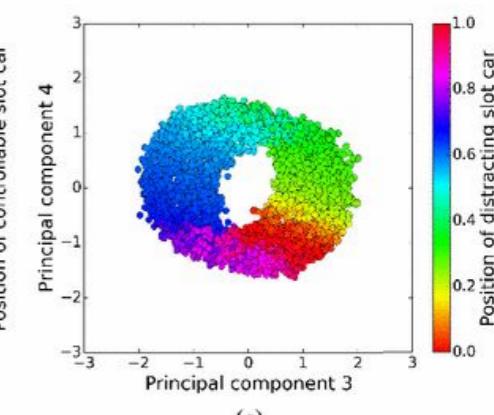
(d)



(e)



(b)



(c)

Evaluation environments: Control theory problems from classic RL literature



Acrobot-v1
 Swing up a two-link robot.



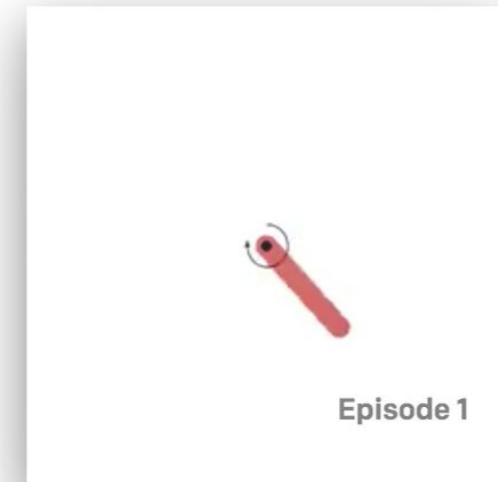
CartPole-v1
 Balance a pole on a cart.



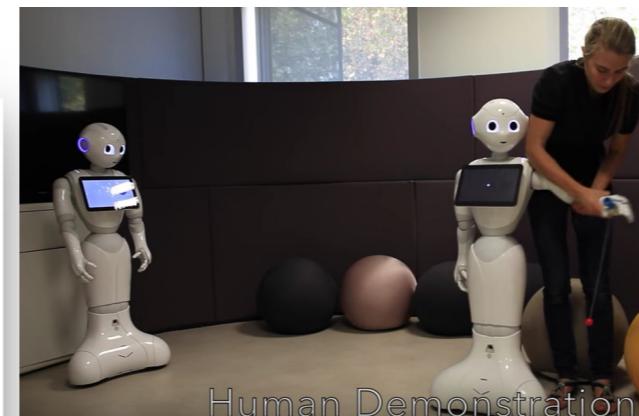
MountainCar-v0
 Drive up a big hill.



MountainCarContinuous-v0
 Drive up a big hill with
continuous control.



Pendulum-v0
 Swing up a pendulum.



Ball in cup



https://gym.openai.com/envs/#classic_control
Octopus arm [Engel'06, Munk'16]

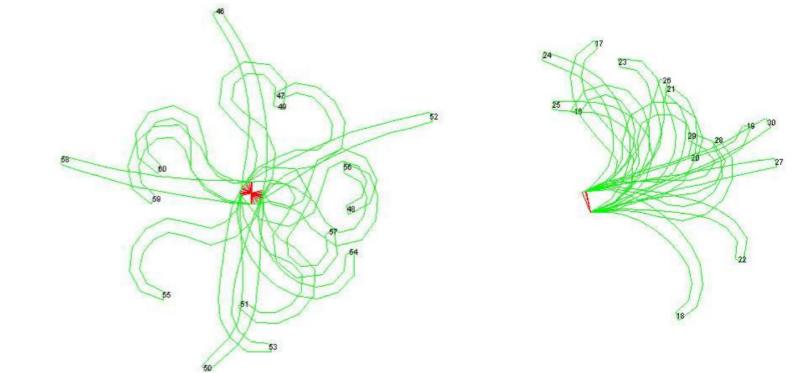


Figure 3: Examples of initial states for the rotating-base experiments (left) and the fixed-base experiments (right). Starting states also include velocities, which are not shown.

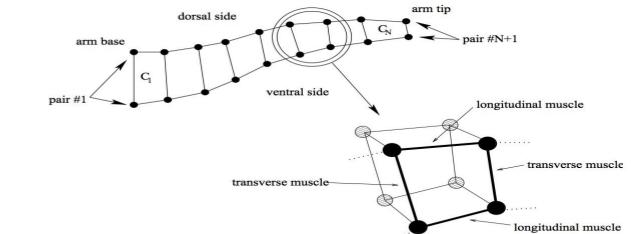


Figure 1: An N compartment simulated Octopus arm. Each constant area compartment C_i is defined by its surrounding 2 longitudinal muscles (ventral and dorsal) and 2 transverse muscles. Circles mark the $2N + 2$ point masses in which the arm's mass is distributed. In the bottom right one compartment is magnified with additional detail.

Figure : Ball-in-cup skill, which the Meka acquired through reinforcement learning



Robotic Priors: Position Velocity Encoders

(PVE) [Jonschkowski et. al. 2017]

$$L_{\text{variation}} = \mathbf{E} \left[e^{-\|\mathbf{s}_a^{(p)} - \mathbf{s}_b^{(p)}\|} \right]$$

$$L_{\text{slowness}} = \mathbf{E} \left[\|\mathbf{s}_t^{(p)} - \mathbf{s}_{t-1}^{(p)}\|^2 \right]$$

$$L_{\text{inertia}} = \mathbf{E} \left[\|\mathbf{s}_t^{(v)} - \mathbf{s}_{t-1}^{(v)}\|^2 \right] = \mathbf{E} \left[\|\mathbf{s}_t^{(a)}\|^2 \right]$$

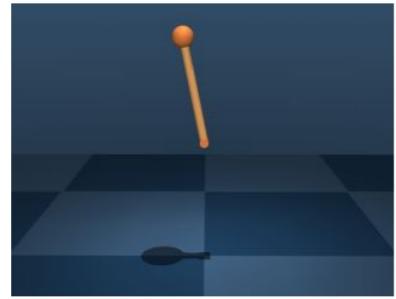
$$L_{\text{conservation}} = \mathbf{E} \left[(\|\mathbf{s}_t^{(v)}\| - \|\mathbf{s}_{t-1}^{(v)}\|)^2 \right]$$

$$L_{\text{controlability (i)}} = e^{-\text{Cov}(\mathbf{a}_{t,i}, \mathbf{s}_{t+1,i}^{(a)})}$$

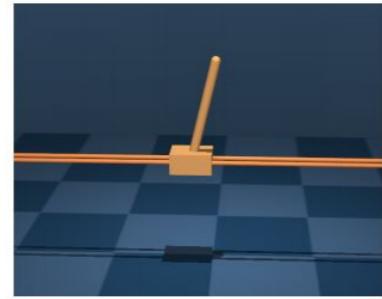
$$= e^{-\mathbf{E} \left[(a_{t,i} - \mathbf{E}[a_{t,i}]) (s_{t+1,i}^{(a)} - \mathbf{E}[s_{t+1,i}^{(a)}]) \right]}$$

Robotic Priors: Position Velocity Encoders

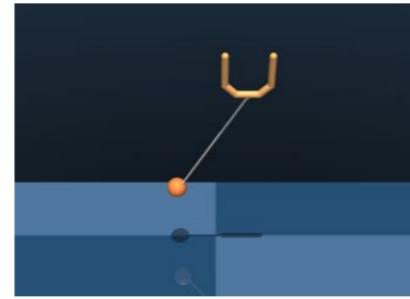
(PVE) [Jonschkowski et. al. 2017]



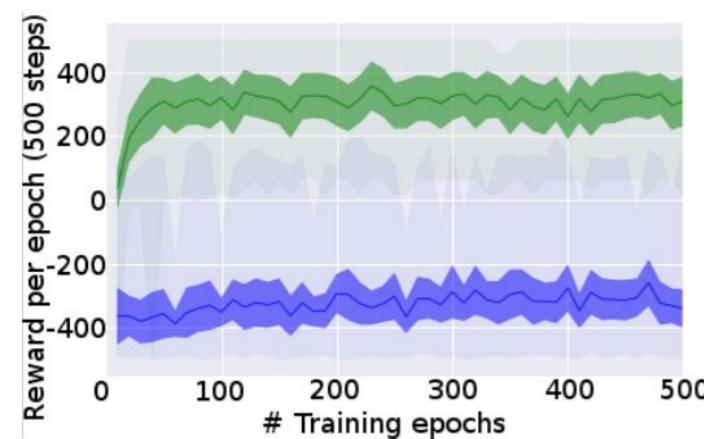
(a) Inverted pendulum



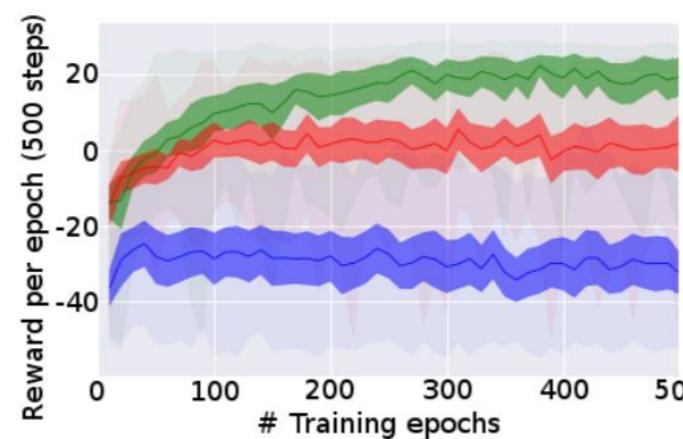
(b) Cart-pole



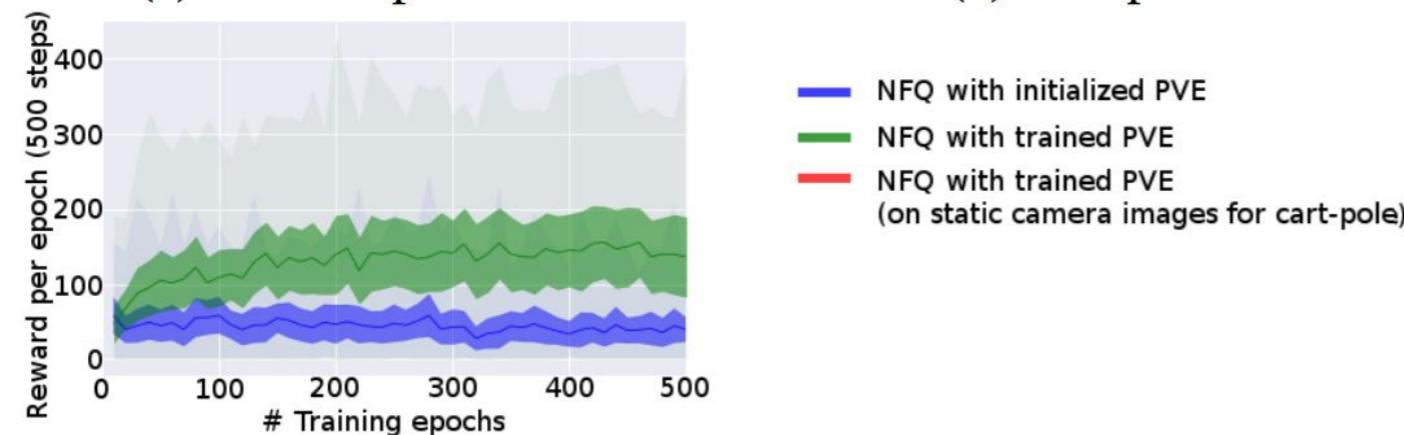
(c) Ball in cup



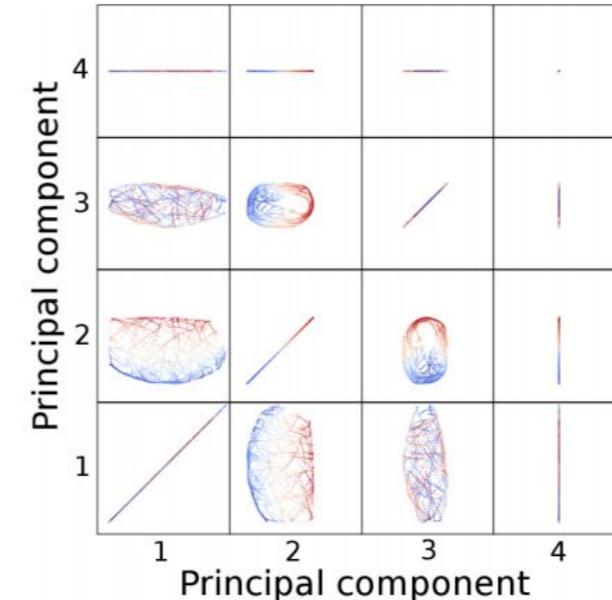
(a) Inverted pendulum



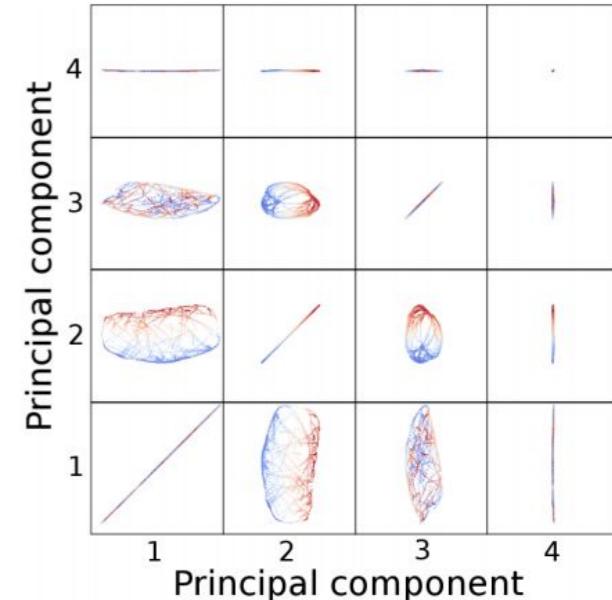
(b) Cart-pole



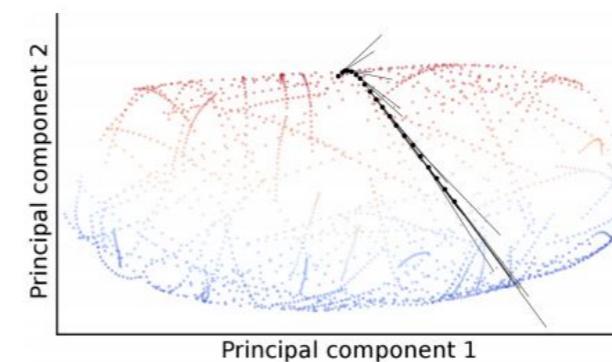
(c) Ball in cup



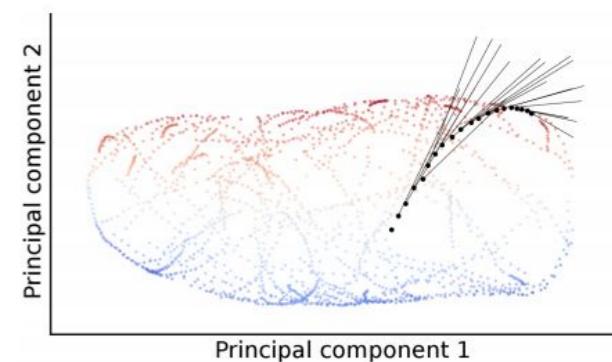
(a) Learned representation
(moving camera)



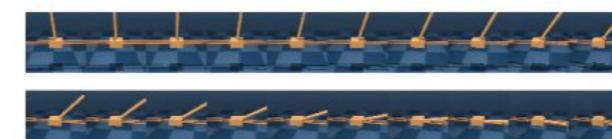
(b) Learned representation
(static camera)



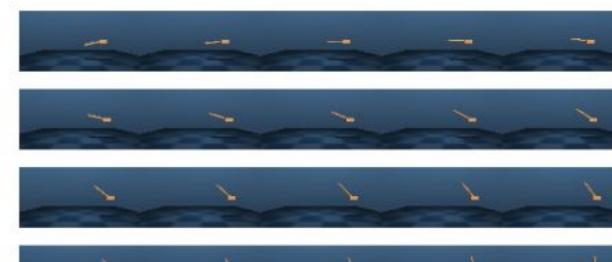
(c) Encoded sequence
(moving camera)



(d) Encoded sequence
(static camera)



(e) Observation sequence
(moving camera)



(f) Observation sequence
(static camera)

BiGAN: Bidirectional Generative Adversarial Network

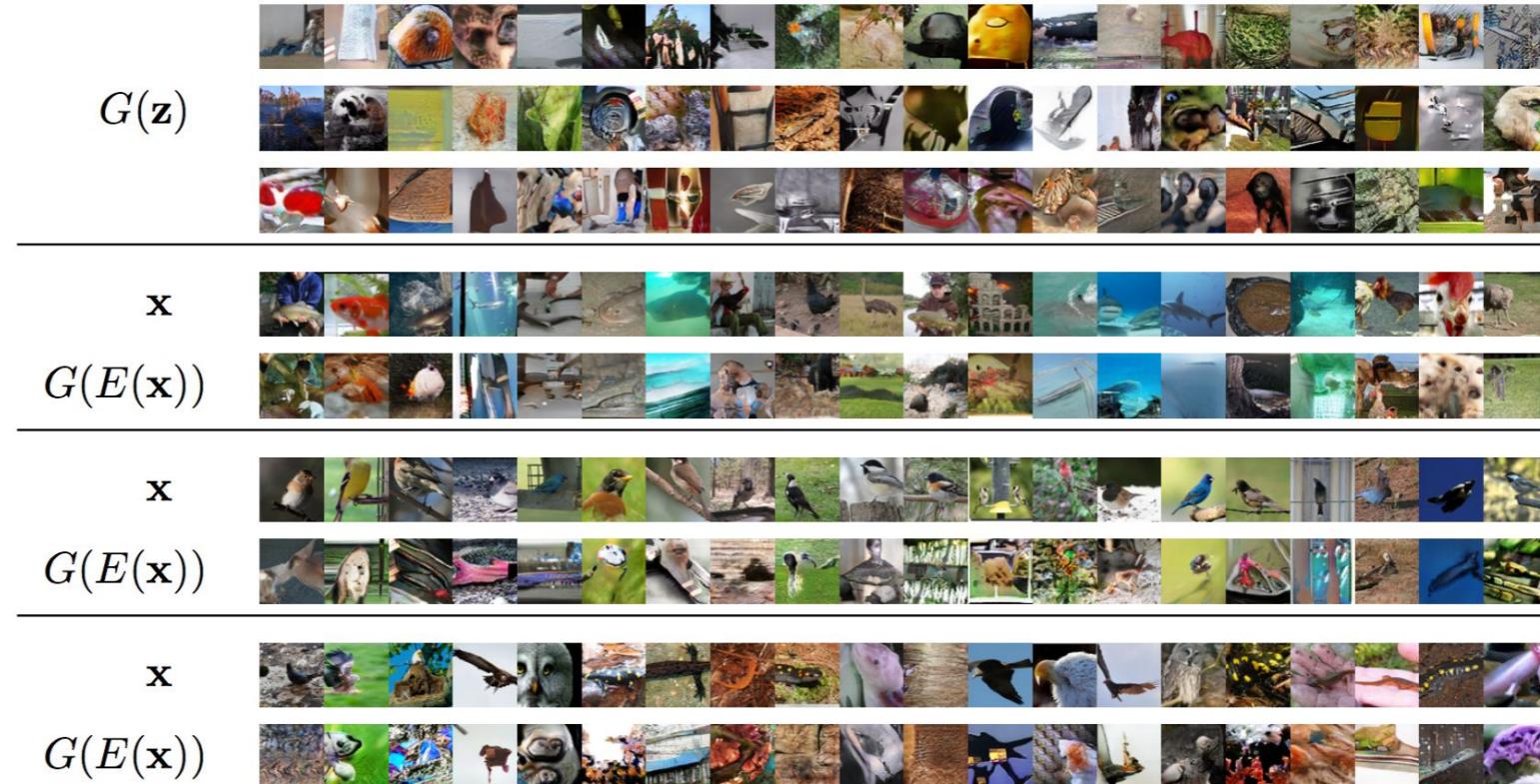
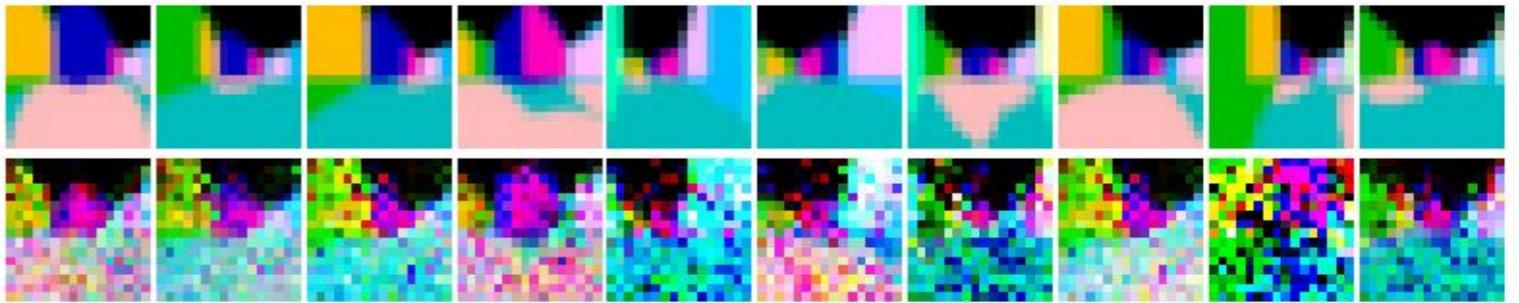


Figure 4: Qualitative results for ImageNet BiGAN training, including generator samples $G(\mathbf{z})$, real data \mathbf{x} , and corresponding reconstructions $G(E(\mathbf{x}))$.

BiGAN for SRL:
observation
reconstruction:

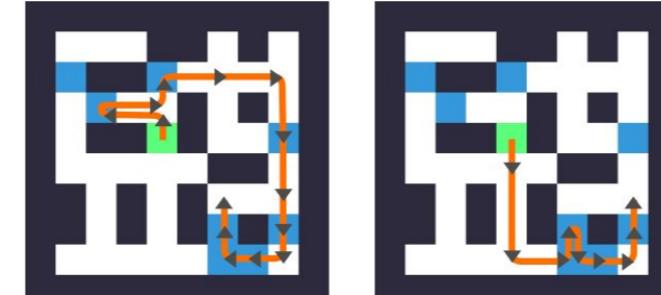


Forward model example: Value Prediction Network [Oh'17]



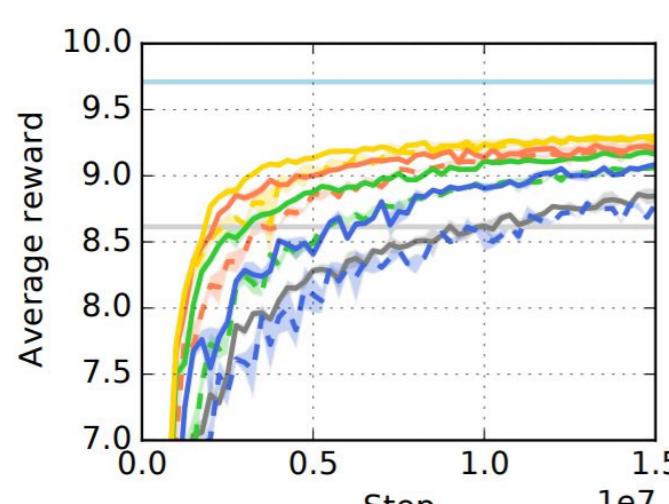
(a) Observation (b) DQN's trajectory (c) VPN's trajectory

Figure 4: Collect domain. (a) The agent should collect as many goals as possible within a time limit which is given as additional input. (b-c) DQN collects 5 goals given 20 steps, while VPN(5) found the optimal trajectory via planning which collects 6 goals.

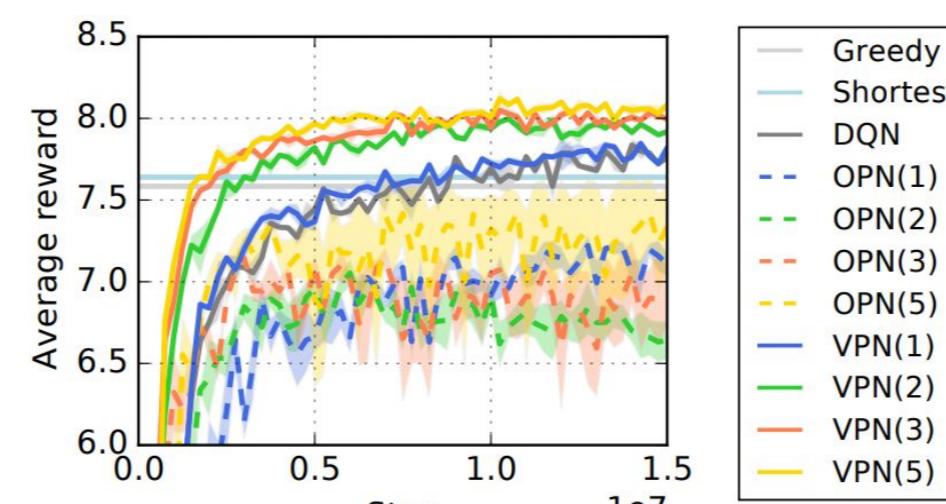


(a) Plan with 20 steps (b) Plan with 12 steps

Figure 5: Example of VPN's plan. VPN can plan the best future options just from the current state. The figures show VPN's different plans depending on the time limit.

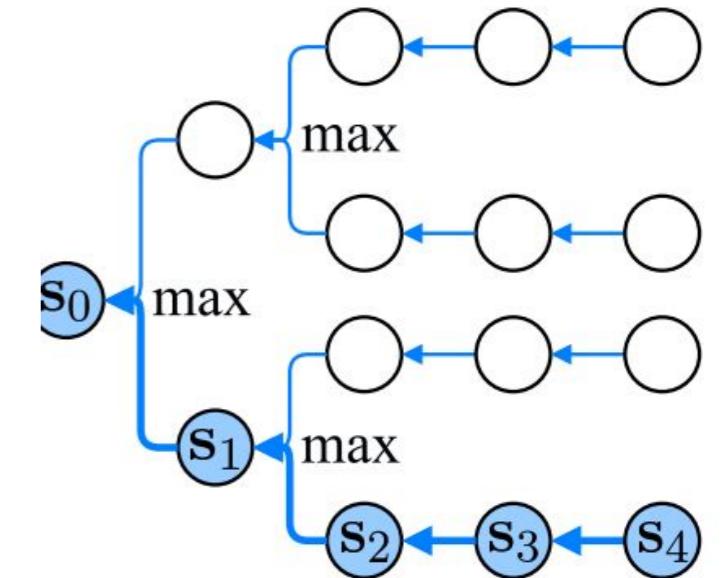


(a) Deterministic



(b) Stochastic

Figure 6: Learning curves on Collect domain. ‘VPN(d)’ represents VPN with d-step planning, while ‘DQN’ and ‘OPN(d)’ are the baselines.



Algorithm 1 Q-value from d -step planning

```

function Q-PLAN( $s, o, d$ )
     $r, \gamma, V(s') \leftarrow f_\theta^{core}(s, o)$ 
    if  $d = 1$  then
        return  $r + \gamma V(s')$ 
    end if
     $\mathcal{A} \leftarrow b\text{-best options based on } Q^1(s', o')$ 
    for  $o' \in \mathcal{A}$  do
         $q_{o'} \leftarrow Q\text{-PLAN}(s', o', d - 1)$ 
    end for
    return  $r + \gamma \left[ \frac{1}{d} V(s') + \frac{d-1}{d} \max_{o' \in \mathcal{A}} q_{o'} \right]$ 
end function

```

What are good states?

*States that are neighbours in the ground truth
should be neighbours in the learned state space*

$$\text{KNN-MSE}(I) = \frac{1}{k} \sum_{I' \in KNN(I, k)} \|\phi(I) - \phi(I')\|^2$$

$$GTC_{(i)} = \max_j |\rho_{s,\tilde{s}}(i, j)| \in [0, 1]$$

Metrics: KNN-MSE:

*States that are neighbours in the ground truth
should be neighbours in the learned state space*

For an image I , this criterion is computed as follows:

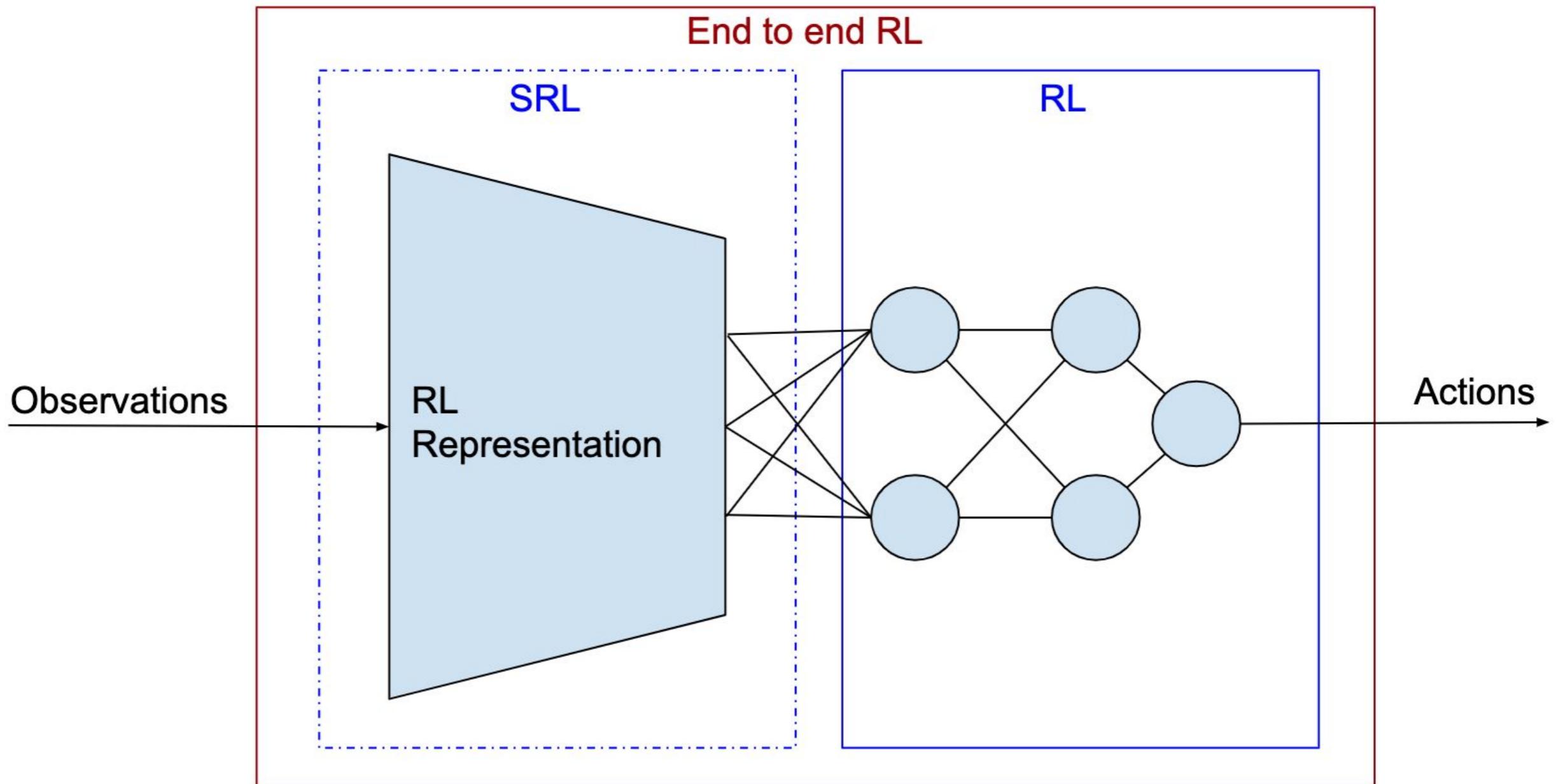
$$\text{KNN-MSE}(I) = \frac{1}{k} \sum_{I' \in \text{KNN}(I, k)} \|\phi(I) - \phi(I')\|^2 \quad (6)$$

where $\text{KNN}(I, k)$ returns the k nearest neighbors of I in the learned state space and $\phi(I)$ gives the ground truth (s) associated to I .

Criterion	GT	Superv	4Priors	5Priors	AE	5Priors 0f
KNN-MSE	0.024	0.03	0.079	0.053	0.099	0.047
NIEQA local	0	0.239	0.66	0.50	0.599	0.52
NIEQA global	0	0.048	0.41	0.20	0.465	0.21

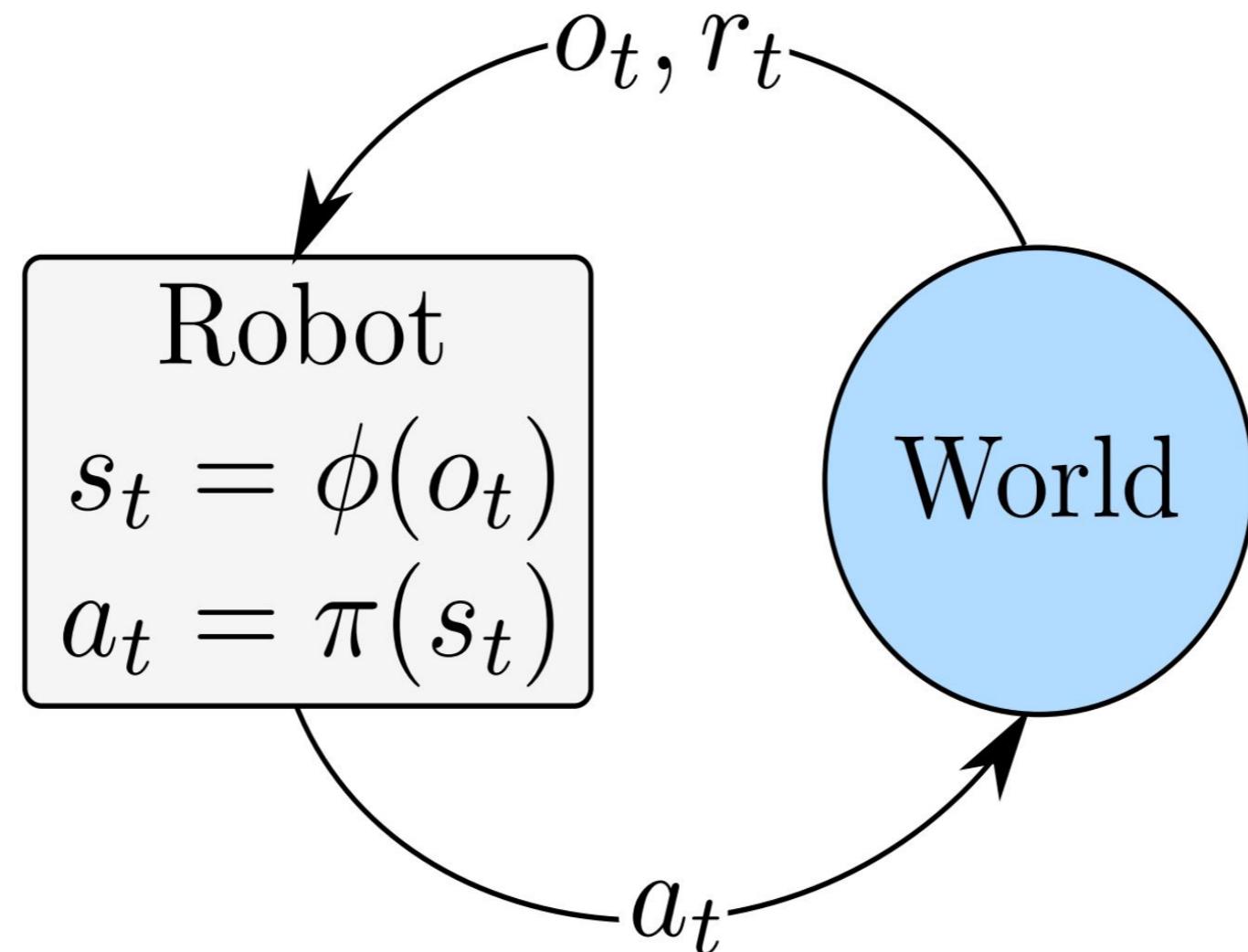
TABLE II
*Static-Button-Distractor 3D (DATASET 2) RESULTS. xF MEANS X RESNET
FROZEN LAYERS, GT: (HAND POSITION) GROUND TRUTH*

State Representation Learning (SRL) in RL context



[Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics \[Raffin et al. 2018\]](#)

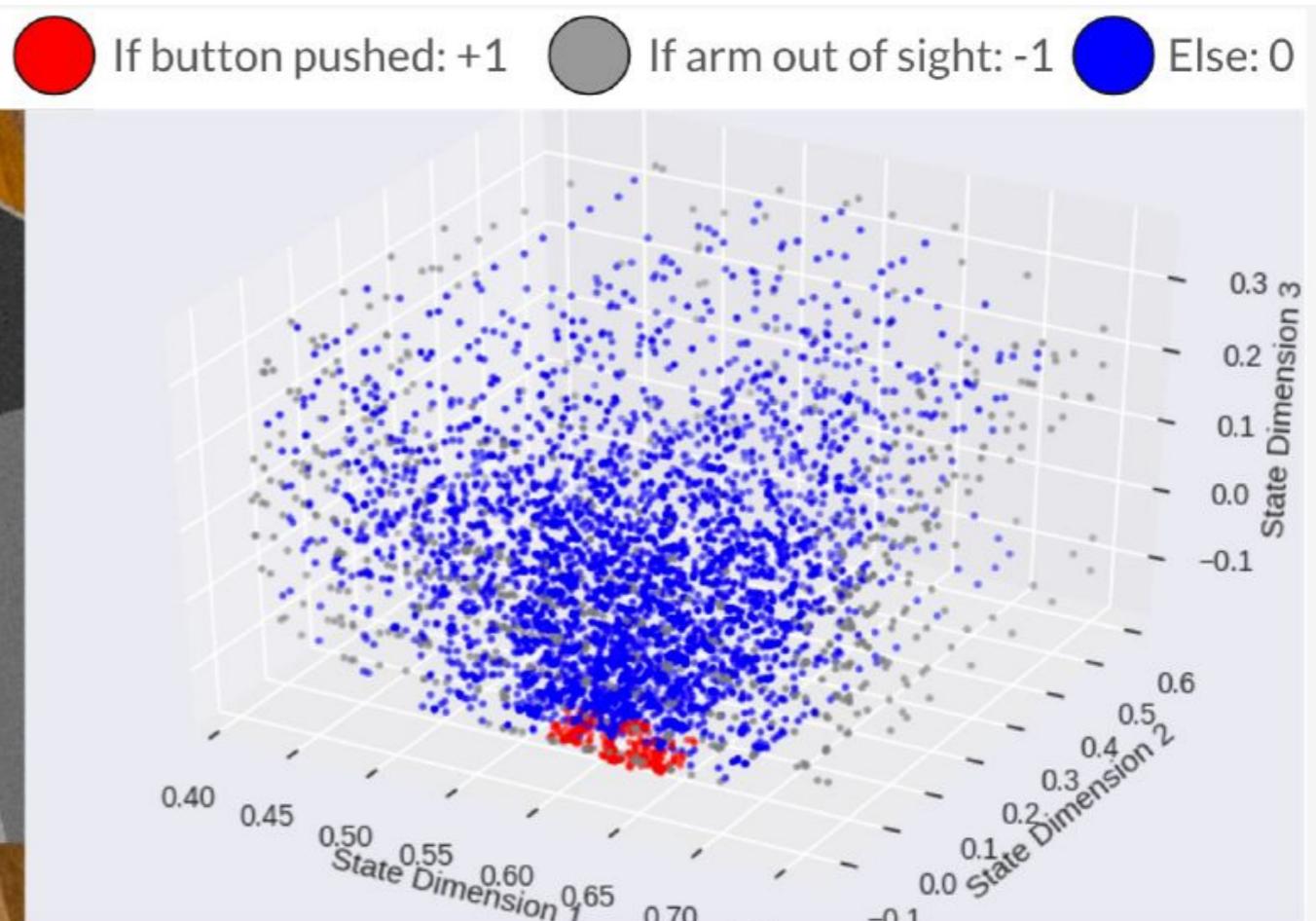
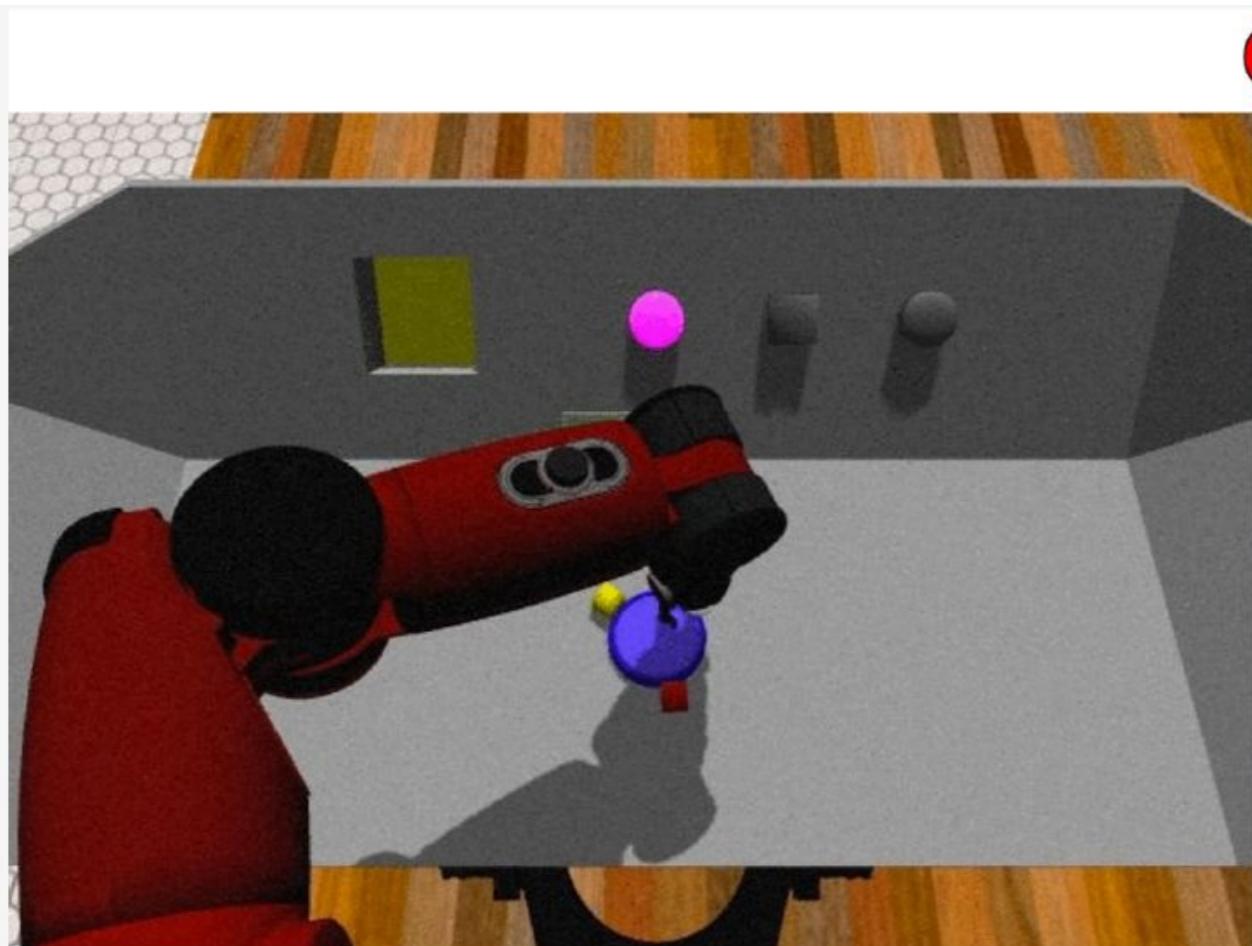
State Representation Learning (SRL) in RL context



SRL Challenges in Robotics



- Vision goal-based tasks: a mobile, visible goal to reach
 - Can't be solved using robot states alone
 - Should encode goal position



What can SRL be used for? e.g., in computer vision for robotics control

Our goal is to learn a relevant state from images, actions and rewards. We train a neural network in unsupervised manner using prior knowledge about the physical world in form of robotic priors.

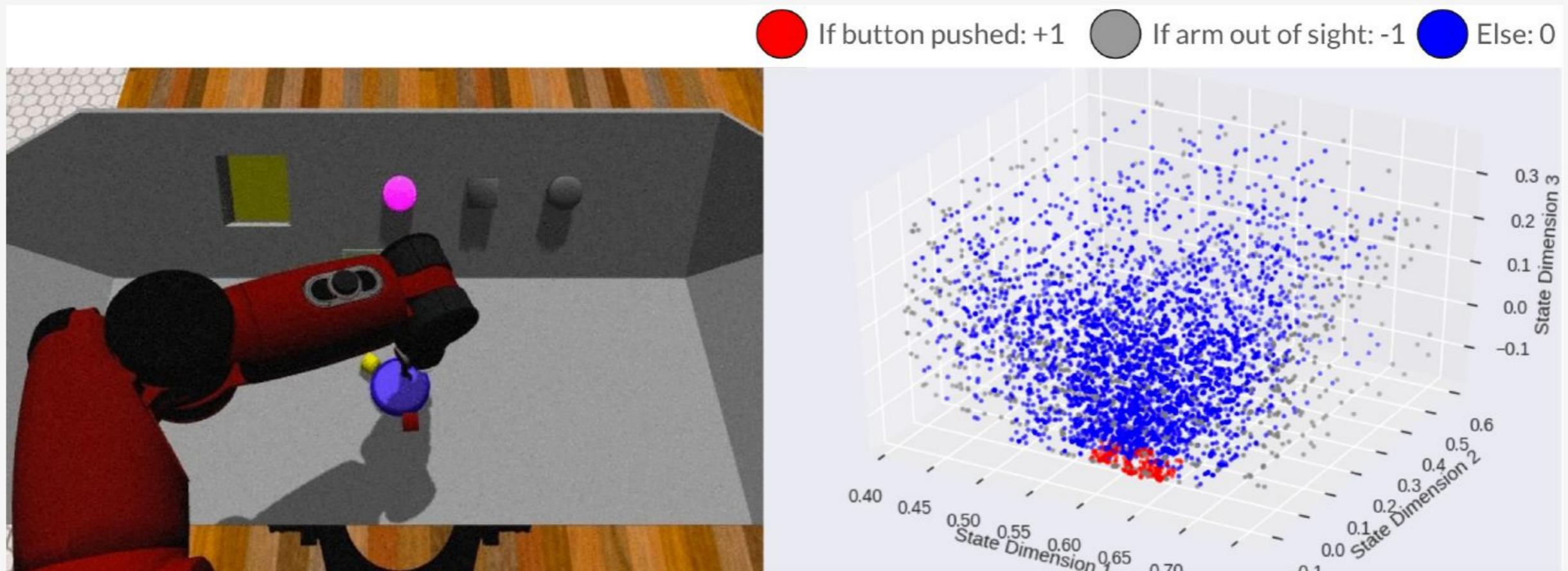


Figure 1: Left: Baxter's camera view for Static-Button-Distractors dataset 2. Right: Baxter's left hand ground truth position and its coded reward