Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics

Antonin Raffin, Ashley Hill, René Traoré, Timothée Lesort, Natalia Díaz, David Filliat
{firstname.lastname}@ensta-paristech.fr
U2IS, ENSTA ParisTech, Université Paris Saclay, Palaiseau, France. Autonomous systems and Robotics http://asr.ensta-paristech.fr/, Inria FLOWERS team https://flowers.inria.fr/

State Representation Learning

Scaling end-to-end reinforcement learning to control real robots from vision presents a series of challenges, in particular in terms of sample efficiency. Against end-to-end learning, state representation learning (SRL) can help learn a compact, efficient and relevant representation of states that speeds up policy learning, reduces the number of samples needed, and is easier to interpret [1]. We evaluate several SRL methods on goal based robotics tasks and propose *SRL Split*, a new unsupervised model that stacks representations and combines strengths of several of these approaches. This method encodes all the relevant features, performs on par or better than end-to-end learning with better sample efficiency, and is robust to hyper-parameters change.

Our SRL model

Using RL notation, SRL corresponds to learning a transformation ϕ from observation o_t to state s_t. Then we learn a policy π that takes state s_t as input and outputs action a_t

$$o_t \xrightarrow{\phi} s_t \xrightarrow{\pi} a_t$$

Our *SRL Splits* model combines a reconstruction of an image I, a reward (r) prediction and an inverse dynamic models losses, using two splits of the state representation s.

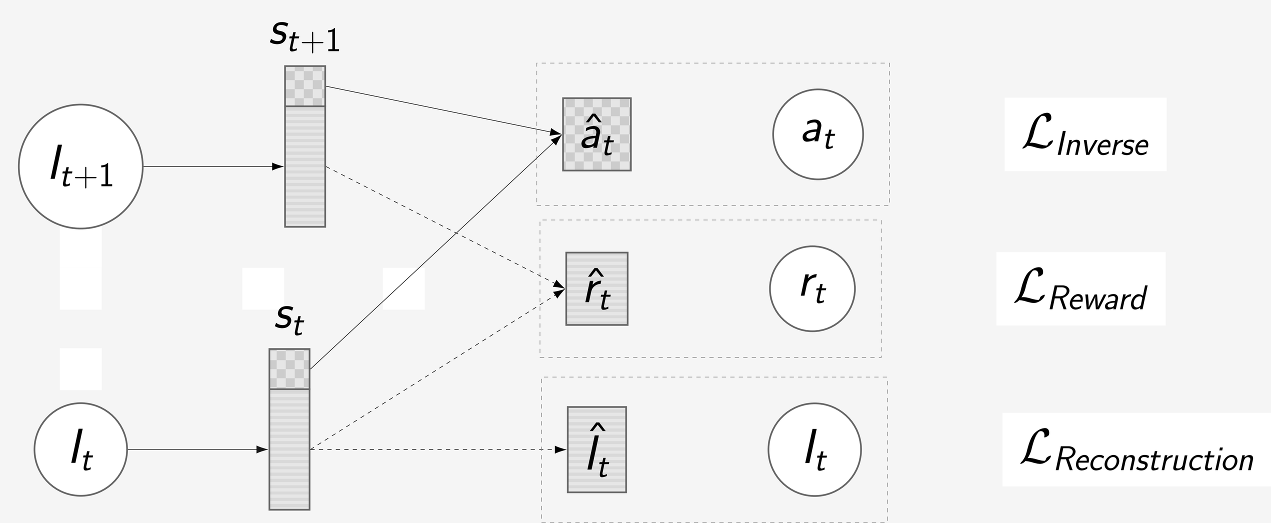
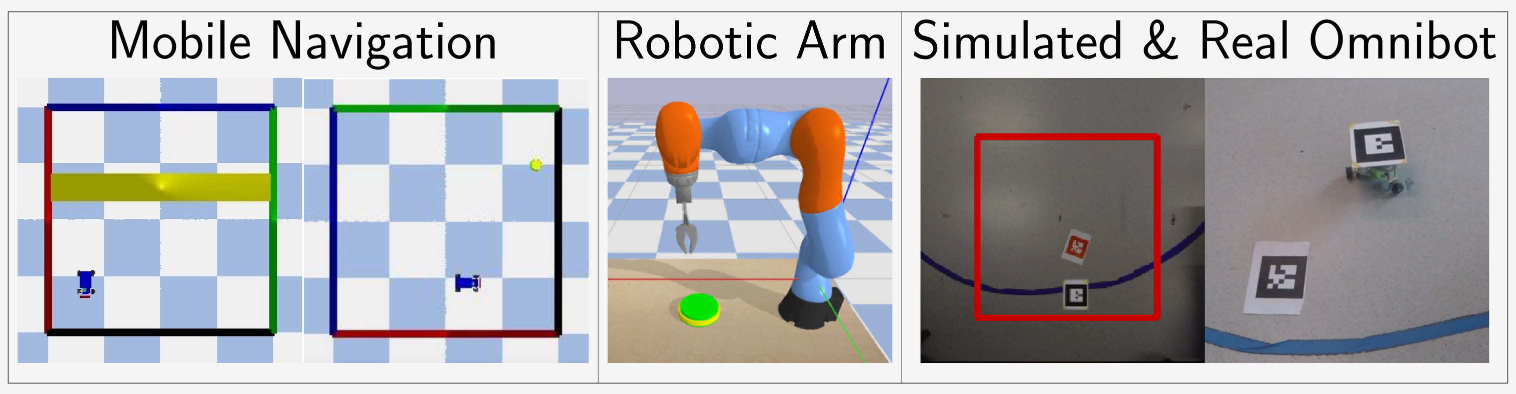


Figure 1: *SRL Splits* model: arrows represent model learning and inference, dashed frames represent losses computation, rectangles are state representations, circles are real observed data, and squares are model predictions.

SRL Datasets and Environments

A set of environments from S-RL Toolbox [2] with variable difficulty was used to assess SRL models covering basic goal-based robotics tasks: mobile navigation and reaching a 3D position.



Quantitative Evaluation

We use the Ground Truth Correlation (GTC) metric [2] that allows to compare the model's ability to encode relevant information:

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

with $i \in [\![0, |\tilde{s}|]\!]$, $j \in [\![0, |s|]\!]$, $\tilde{s} = [\tilde{s}_1; \ldots; \tilde{s}_n]$, and $\tilde{s}_k$ being the k^{th} dimension of the ground truth state vector. The mean of GTC allows to compare learned states using one scalar value: $GTC_{mean} = \mathbb{E}[GTC]$.

References

[1] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat.
State representation learning for control: An overview.
Neural Networks, 2018.

[2] A. Raffin, A. Hill, R. Traoré, T. Lesort, N. Díaz-Rodríguez, and D. Filliat.
S-RL toolbox: Environments, datasets and evaluation metrics for state representation learning.
In *NeurIPS Workshop on Deep Reinforcement Learning*, 2018.

Experiments

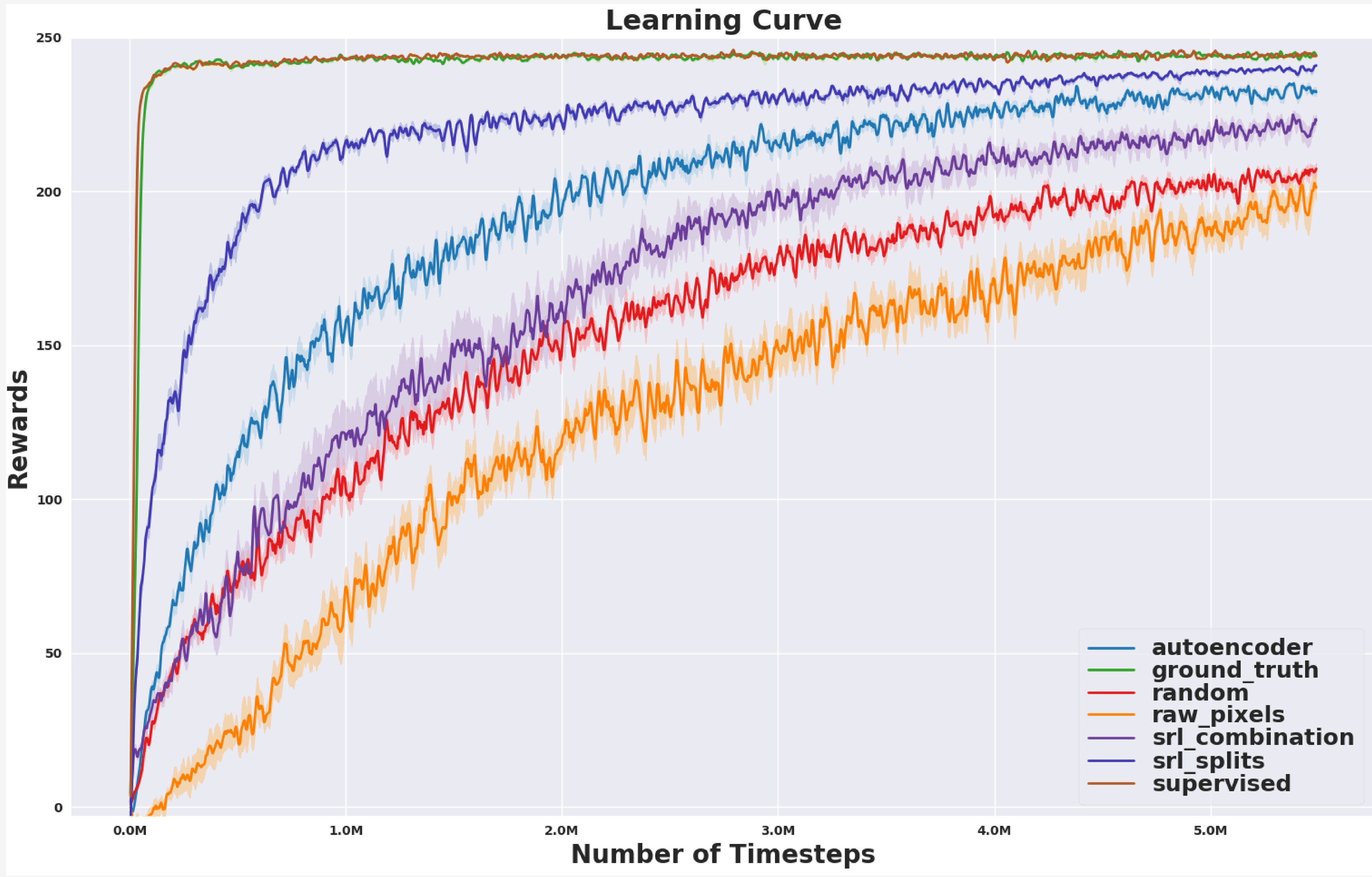


Figure 2: Performance (mean and standard error for 8 runs) for PPO algorithm for different state representations learned in Simulated Omnibot with randomly initialized target environment.

Ground Truth Correlation	x_{robot}	y_{robot}	x_{target}	y_{target}	*Mean*	*Mean Reward*
Ground Truth	1	1	1	1	1	243.7 ± 1.2
Supervised	0.69	0.73	0.6	0.61	0.66	243.9 ± 1.8
Random Features	0.59	0.54	0.50	0.42	0.51	201.5 ± 5.7
Robotic Priors	0.1	0.1	0.45	0.54	0.30	-1.1 ± 2.4
Auto-Encoder	0.50	0.54	0.20	0.25	0.37	230.27 ± 3.2
SRL Combination	0.95	0.96	0.22	0.20	0.58	216.8 ± 5.6
SRL Splits	0.98	0.98	0.61	0.73	0.83	237.8 ± 2.1

Table 1: GTC, GTC_{mean}, and mean reward performance in RL (using PPO) per episode after 5 millions steps, with standard error (SE) for each SRL method in 2D Simulated Omnibot with a random target environment.

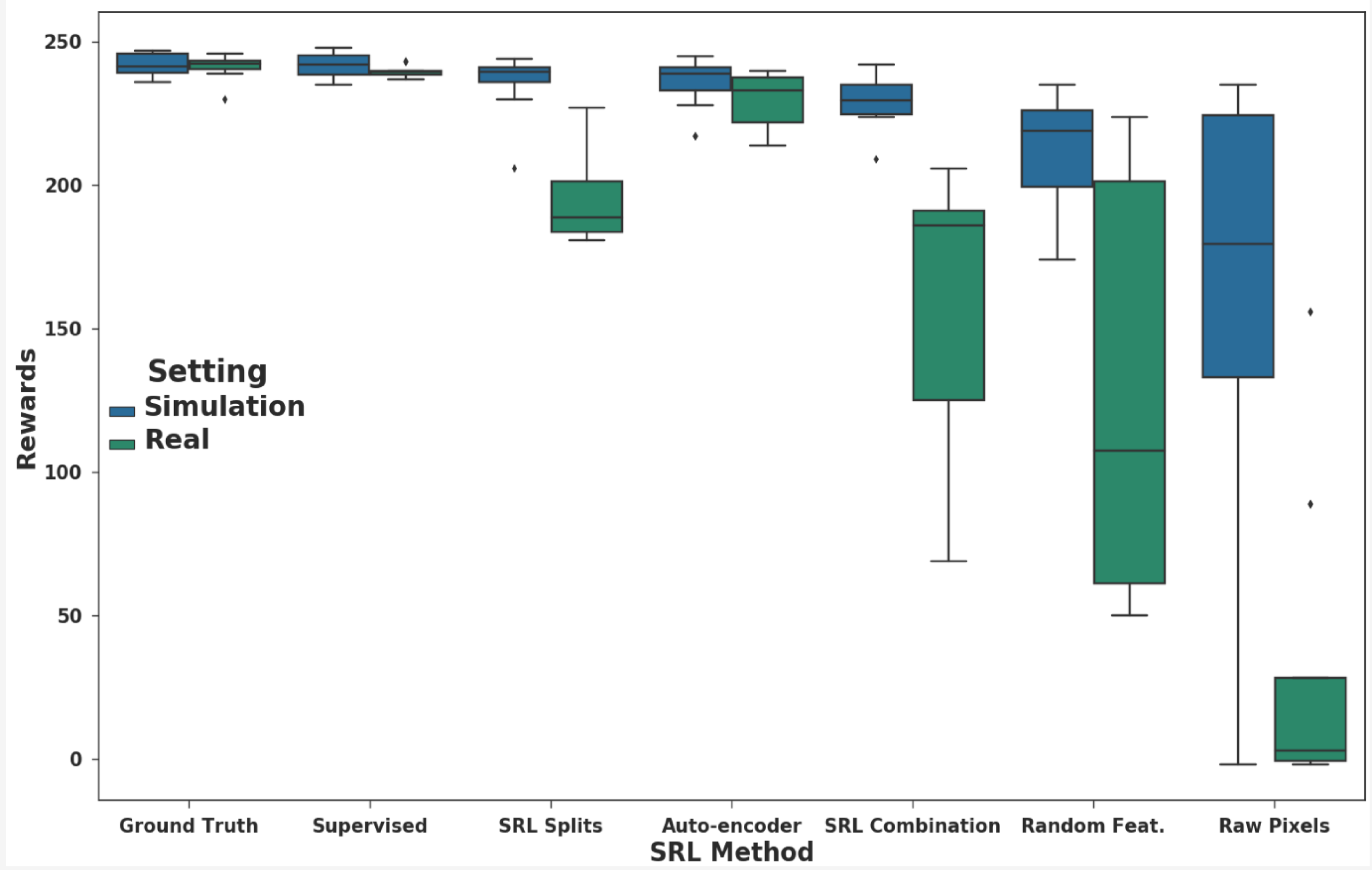


Figure 3: **From simulation to real robot**: Mean reward and standard deviation for policies trained in simulation (5M steps budget) and replayed in Simulated and Real Omnibot (250 steps, 8 runs).

Conclusion

We show the advantages of decoupling feature extraction from policy learning in RL on a set of goal-based robotics tasks. We also show that random features are a good baseline versus end-to-end learning, and introduce the *SRL Splits* model, which is robust against perturbations and helps transfer to a real robot.

- **Repository**: https://github.com/araffin/srl-zoo
- **Acknowledgement**: This work is supported by the EU H2020 DREAM project (Grant agreement No 640891)