



Unsupervised State Representation Learning in Partially Observable Atari Games

Li Meng¹(✉) , Morten Goodwin^{2,4} , Anis Yazidi³ , and Paal Engelstad¹ 

¹ University of Oslo, Oslo, Norway

{li.meng,paal.engelstad}@its.uio.no

² Centre for Artificial Intelligence Research, University of Agder, Kristiansand, Norway

morten.goodwin@uia.no

³ Oslo Metropolitan University, Oslo, Norway

anisy@oslomet.no

⁴ Oslo Metropolitan University, Kristiansand, Norway

Abstract. State representation learning aims to capture latent factors of an environment. Although some researchers realize the connections between masked image modeling and contrastive representation learning, the effort is focused on using masks as an augmentation technique to represent the latent generative factors better. Partially observable environments in reinforcement learning have not yet been carefully studied using unsupervised state representation learning methods.

In this article, we create an unsupervised state representation learning scheme for partially observable states. We conducted our experiment on a previous Atari 2600 framework designed to evaluate representation learning models. A contrastive method called Spatiotemporal DeepInfomax (ST-DIM) has shown state-of-the-art performance on this benchmark but remains inferior to its supervised counterpart. Our approach improves ST-DIM when the environment is not fully observable and achieves higher F1 scores and accuracy scores than the supervised learning counterpart. The mean accuracy score averaged over categories of our approach is $\sim 66\%$, compared to $\sim 38\%$ of supervised learning. The mean F1 score is $\sim 64\%$ to $\sim 33\%$. The code can be found on https://github.com/mengli11235/MST_DIM.

Keywords: State representation Learning · Contrastive learning

1 Introduction

Deep representation learning is a machine learning (ML) type that focuses on learning useful data representations. These representations can be learned using deep neural networks (NNs) and transferred to a variety of downstream computer vision (CV), and natural language processing (NLP) tasks [7, 15]. Deep representation learning includes autoencoders [14], generative models [9], contrastive methods [11, 20] and transformer models [6].

State representation learning (SRL) [1, 13, 18] is a particular field of representation learning where the state observations are commonly seen in the reinforcement learning (RL) setup. Agents can interact with the environment, which itself changes accordingly throughout interactions. RL is a well-established ML field that solves the Markov decision process (MDP) [21]. Traditional RL methods such as Q-learning [26] have evolved through adopting NNs [19]. Moreover, convolutional neural networks (CNNs) are deployed in environments with image inputs, and RL agents can learn from raw pixels.

Partially observable Markov decision processes (PODMPs) [3, 23] are MDPs where an agent can only observe a limited part of the environment, not the full state. Recently, there have been some developments in decoupling SRL from RL [10, 16, 17, 22]. More improvements by using SRL have been reported in partially observable environments than in fully observable ones. However, it is not clear how the POMDP state is captured and preserved by representations.

This paper designs an unsupervised representation learning scheme for partially observable environments. This method extends ST-DIM [1] and introduces an unsupervised pretraining setting suitable to partially observable Atari Games. Different pretraining hyper-parameter choices are also discussed in our ablation study. Our contribution is summarized as follows: (1) We propose MST-DIM, a contrastive method suitable to pretrain data collected in a partially observable environment. (2) We test our method on the SRL benchmark using 20 Atari 2600 games and compare the results with the ST-DIM and supervised methods. (3) Should the percentages of the observable parts of states be the same in pre-training and in probing? Extensive evaluations are conducted to examine what is needed for SRL in order to let the model accurately predict the ground truth labels.

2 Related Work

Self-supervised Learning. Self-supervised learning learns useful representations from unlabeled data, which can be used in various downstream tasks. The methodology has played an important role in NLP [6] and CV fields. Contrastive Predictive Coding (CPC) [20] learns predictive representations by capturing the information that is maximally useful to predict future (spatial or temporal) samples. SimCLR [4] provides a simple yet effective framework for contrastive learning. Momentum Contrast (MoCo) keeps dynamic dictionaries for contrastive learning [11]. Self-supervised Vision Transformers [5] study the usage of ViTs ([8]) on MoCo. SRL benefits from those paradigms by having a low dimensional state space that can learn a control policy more efficiently [18].

Contrastive Representations for RL. Contrastive Unsupervised Representations for RL (CURL) [16] is an RL pipeline that extracts high-level features using an auxiliary contrastive loss, which can be combined with on-policy or off-policy RL algorithms. Masked Contrastive Representation Learning (M-CURL) [27] improves the data efficiency in CURL by considering the correlation among consecutive inputs and using masks to help the transformer module learn to

reconstruct the features of the ground truth. The loss is defined as a sum of the RL loss and masked contrastive loss, and the transformer module is discarded during inference. M-CURL reportedly outperforms CURL on 21 out of 26 environments from Atari 2600 Games.

Representation Learning in POMDPs. Predictions of Bootstrapped Latents (PBL) [10] is an RL algorithm designed for the multitask setting. PBL is trained by predicting future latent observations from partial histories and the current states from latent observations. Histories in POMDPs are typically compressed into a current agent state using NNs [23]. Agents trained with PBL achieve significantly higher human normalized scores than baseline methods in the partially observable DMLab 30 environments, but the gap is at most minimal in fully observable environments. Augmented Temporal Contrast (ATC) [22] associates temporally close pairs of observations and also shows its usefulness in POMDPs.

3 Method

ST-DIM [1] is a method that captures the latent generative factors through maximizing mutual information lower-bound estimate over consecutive observations x_t and x_{t+1} given a set of cross-episode observations $\chi = \{x_1, x_2, \dots, x_n\}$, originated from agents interacting with RL environments. It uses infoNCE [20] that maximizes Eq. 1 as the mutual information estimator between patches as Deep InfoMax (DIM) [12] does.

$$\mathcal{I}_{NCE}(\{(x_i, y_i)\}_{i=1}^N) = \sum_{i=1}^N \log \frac{\exp f(x_i, y_i)}{\sum_{j=1}^N \exp f(x_i, y_j)} \quad (1)$$

For any i , (x_i, y_i) is called positive examples from the joint distribution $p(x, y)$ and (x_i, y_j) from the product of marginals $p(x)p(y)$ is called negative examples for any $i \neq j$. Meanwhile, $f(x, y)$ is a score function, i.e., a bilinear layer.

ST-DIM utilizes both the global-local (Eq. 2) and local-local objective (Eq. 3). An illustration of the global-local contrastive task is also shown in Fig. 1. The difference between the local-local and global-local tasks is that an additional MLP is used to extract the global features.

$$\mathcal{L}_{GL} = \sum_{m=1}^M \sum_{n=1}^N -\log \frac{\exp g_{m,n}(x_t, x_{t+1})}{\sum_{x_{t*} \in X_{n \in x_t}} \exp g_{m,n}(x_t, x_{t*})} \quad (2)$$

$$\mathcal{L}_{LL} = \sum_{m=1}^M \sum_{n=1}^N -\log \frac{\exp f_{m,n}(x_t, x_{t+1})}{\sum_{x_{t*} \in X_{n \in x_t}} \exp f_{m,n}(x_t, x_{t*})} \quad (3)$$

Here, M and N are the height and width, $g_{m,n}(x_t, x_{t+1}) = \phi(x_t)^T W_g \phi_{m,n}(x_{t+1})$ and $\phi(m, n)$ is the local feature vector produced by convolutional layers in the representation encoder ϕ at the location (m, n) . On the other hand, $f_{m,n}(x_t, x_{t+1}) = \phi_{m,n}(x_t)^T W_l \phi_{m,n}(x_{t+1})$. Observations x_t and x_{t+1}

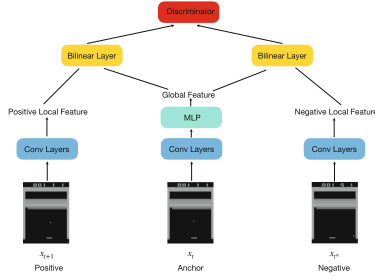


Fig. 1. An illustration of the global-local contrastive task in ST-DIM. For the local-local contrastive task, we discard the MLP and use the local feature of the anchor.

are temporally adjacent, whereas x_{t^*} is a randomly sampled observation from the minibatch.

In order to fit ST-DIM into partially observable environments, we propose MST-DIM and define random masks k_t for each consecutive pair (x_t, x_{t+1}) that is drawn from a binomial distribution \mathbb{K} . Therefore, Eq. 2 and Eq. 3 are modified as Eq. 4 and Eq. 5:

$$\mathcal{L}_{MGL} = \sum_{m=1}^M \sum_{n=1}^N -\log \frac{\exp g_{m,n}(x_t k_t, x_{t+1})}{\sum_{x_{t^*} \in X_{n \text{ ext}}} \exp g_{m,n}(x_t k_t, x_{t^*})} \quad (4)$$

$$\mathcal{L}_{MLL} = \sum_{m=1}^M \sum_{n=1}^N -\log \frac{\exp f_{m,n}(x_t k_t, x_{t+1})}{\sum_{x_{t^*} \in X_{n \text{ ext}}} \exp f_{m,n}(x_t k_t, x_{t^*})} \quad (5)$$

Meanwhile, we can tune the probability of masking in \mathbb{K} for both the pretraining and probing. For example, a masking ratio of 0.4 means that 40% of the full observation is not visible to the agent. As a result, our method not only allows pretraining in partially observable environments, but is also capable of pretraining even though the dataset comes from a different \mathbb{K} distribution.

4 Experimental Details

Our experiment is conducted among 20 Atari games of Arcade Learning Environment (ALE). Performances are evaluated by probe accuracy and F1 scores for each game. Because ALE does not directly provide ground truth information, ST-DIM has been conducted on the newly designed Atari Annotated RAM Interface (AtariARI) [1] that exposes the state variables from the source code [24] in 22 games. State variables are categorized as agent localization (Agent Loc.), small object localization (Small Loc.), other localization (Other Loc.), score/clock/lives/display (Score/.../Display), and miscellaneous (Misc.). Detailed descriptions of states for each game across categories can be found in the original paper [1]. We also summarize probe accuracy and F1 scores across

those state categories for each game in our experiment. Not all categories are available for each game, and we only include results from applicable ones.

Due to practical implementation issues, we exclude Berzerk, Riverraid and Yars Revenge and include Battle Zone in our experiment, making a total number of 20 games. Trajectories collected by random agents are used in our experiment, as ST-DIM [1] suggested it can be a better choice than using trajectories from PPO agents.

For pretraining, we use different partially observable setups. To verify if the masking ratio in pretraining can be different from the probing, five types of pretraining images have been considered in our experiment, as illustrated by Fig. 2. The images can be original, 20% masked, 40% masked, 60% masked, or 80% masked.

We follow the same probing protocol as ST-DIM and focus on the explicitness, i.e., how well the latent generative factors can be recovered. This is done by training a linear classifier that predicts the state variables using the learned representations. We keep the hyper-parameters the same as ST-DIM to make our experiment comparable. A short list of hyper-parameters is shown in Table 1. Setting the entropy threshold removes large objects that have low entropy from the labels. The encoder architecture is the same as in ST-DIM, as illustrated by Fig. 3.

Table 1. Parameter Choices

Hyper-parameter	Value
Image Size	160×210
Batch Size	64
Learning Rate	$3e-4$
Entropy Threshold	0.6
Pretraining Steps	80000
Probe Training Steps	35000
Probe Testing Steps	10000



Fig. 2. The original image is masked by different percentages of random noise.

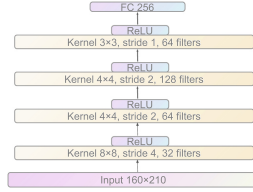


Fig. 3. Architecture of the encoder.

Table 2. Probe F1 scores of each game averaged across categories

Games	Observable	Non-obs.	Supervised	Pretrain	Ratio 0.2	Ratio 0.6	Ratio 0.8
Asteroids	0.46	0.39	0.39	0.45	0.45	0.44	0.44
Battle Zone	0.5	0.29	0.28	0.45	0.39	0.41	0.38
Bowling	0.96	0.29	0.29	0.9	0.72	0.85	0.63
Boxing	0.59	0.11	0.1	0.53	0.38	0.43	0.2
Breakout	0.87	0.37	0.37	0.85	0.83	0.85	0.29
Demon Attack	0.66	0.46	0.45	0.64	0.6	0.63	0.57
Freeway	0.81	0.03	0.03	0.27	0.27	0.1	0.05
Frostbite	0.72	0.33	0.34	0.7	0.65	0.66	0.58
Hero	0.92	0.57	0.58	0.9	0.87	0.88	0.84
Ms Pacman	0.7	0.35	0.36	0.69	0.64	0.68	0.62
Montezuma Revenge	0.77	0.54	0.53	0.75	0.74	0.74	0.71
Pitfall	0.68	0.24	0.25	0.62	0.53	0.6	0.54
Pong	0.81	0.13	0.13	0.71	0.69	0.65	0.42
Private Eye	0.88	0.5	0.49	0.84	0.81	0.82	0.68
Qbert	0.72	0.47	0.47	0.71	0.69	0.71	0.69
Seaquest	0.64	0.38	0.37	0.62	0.59	0.61	0.54
Space Invaders	0.56	0.45	0.44	0.56	0.55	0.57	0.56
Tennis	0.6	0.13	0.13	0.48	0.34	0.4	0.23
Venture	0.55	0.39	0.4	0.54	0.53	0.53	0.52
Video Pinball	0.62	0.29	0.3	0.62	0.57	0.63	0.6
Mean	0.7	0.34	0.33	0.64	0.59	0.61	0.5

5 Results

In this section, we demonstrate the performance of unsupervised representation learning in the partially observable reinforcement learning environment. Table 2 shows the F1 scores for each game and Table 3 shows the accuracy. “Observable” represents the setting where the probing is implemented with full observations. “Non-obs.” is the ST-DIM setting where the pretraining is with full observations, but the probing are with partial observations. “Supervised” is the setting where no pretraining is included but the model is trained and tested only in probing using the supervised manner. “Ratio” indicates the masking ratio in probing,

Table 3. Probe accuracy scores of each game averaged across categories

Games	Observable	Non-obs.	Supervised	Pretrain	Ratio 0.2	Ratio 0.6	Ratio 0.8
Asteroids	0.5	0.46	0.46	0.5	0.5	0.5	0.49
Battle Zone	0.52	0.36	0.35	0.48	0.44	0.45	0.44
Bowling	0.96	0.36	0.36	0.91	0.73	0.85	0.66
Boxing	0.59	0.14	0.13	0.54	0.39	0.44	0.22
Breakout	0.88	0.41	0.41	0.86	0.84	0.86	0.37
Demon Attack	0.66	0.47	0.47	0.65	0.6	0.64	0.58
Freeway	0.81	0.06	0.06	0.3	0.3	0.15	0.09
Frostbite	0.73	0.38	0.38	0.7	0.66	0.67	0.59
Hero	0.92	0.59	0.59	0.9	0.88	0.88	0.84
Ms Pacman	0.71	0.4	0.4	0.7	0.66	0.69	0.65
Montezuma Revenge	0.77	0.55	0.54	0.76	0.74	0.74	0.72
Pitfall	0.69	0.3	0.3	0.64	0.55	0.62	0.57
Pong	0.82	0.21	0.21	0.73	0.7	0.67	0.47
Private Eye	0.88	0.52	0.51	0.84	0.81	0.83	0.68
Qbert	0.73	0.51	0.51	0.72	0.7	0.71	0.69
Seaquest	0.66	0.46	0.45	0.63	0.61	0.63	0.57
Space Invaders	0.57	0.48	0.47	0.59	0.57	0.59	0.58
Tennis	0.61	0.22	0.22	0.51	0.39	0.44	0.29
Venture	0.56	0.41	0.42	0.55	0.54	0.54	0.53
Video Pinball	0.63	0.31	0.32	0.62	0.57	0.63	0.61
Mean	0.71	0.38	0.38	0.66	0.61	0.63	0.53

e.g., ratio 0.2 equals that 20% part of each observation in probing is not visible. By default, the masking ratio in probing is set to 0.4.

It is obvious that the model is capable of predicting the state variables the most in fully observable environments, as the observations in probing are not masked by noise. The rest results are all from models trained and tested with masked observations for probing. The performance deteriorates considerably if the model still uses ST-DIM and pretrains on full observations. Supervised training also exhibits the same characteristic and obtains similar results with ST-DIM. On the other hand, MST-DIM achieves significantly higher accuracy and F1 scores by taking advantage of masked pretraining. Pretraining with a different masking ratio has also shown improvements over ST-DIM. Using masking ratios 0.2 and 0.6 in pretraining yields slightly inferior results than the default masking ratio of 0.4. Surprisingly, using a masking ratio of 0.8 still improves the ST-DIM despite most of the images being masked in this setting.

The mean accuracy score (0.66) and F1 score (0.64) of MST-DIM are slightly worse than the accuracy score (0.71) and F1 score (0.7) under the fully observable setup. However, these results considerably exceed those of supervised learning (0.38 and 0.33) and ST-DIM (0.38 and 0.34) under the same partially observable setup. For the other three masking ratios, a ratio of 0.6 achieves the highest accuracy (0.63) and F1 scores (0.61), a ratio of 0.2 achieves slightly lower scores

Table 4. Probe F1 scores of different ground truth categories averaged across all games

Categories	Observable	Non-obs.	Supervised	Pretrain	Ratio 0.2	Ratio 0.6	Ratio 0.8
Agent Loc	0.58	0.26	0.26	0.52	0.48	0.5	0.41
Misc.	0.73	0.48	0.48	0.72	0.68	0.7	0.61
Other Loc.	0.64	0.34	0.34	0.59	0.56	0.54	0.47
Score/.../Display	0.9	0.42	0.42	0.86	0.77	0.83	0.7
Small Loc.	0.53	0.21	0.21	0.47	0.42	0.44	0.28

Table 5. Probe accuracy scores of different ground truth categories

Categories	Observable	Non-obs.	Supervised	Pretrain	Ratio 0.2	Ratio 0.6	Ratio 0.8
Agent Loc	0.59	0.32	0.32	0.54	0.5	0.52	0.44
Misc.	0.74	0.52	0.52	0.73	0.69	0.71	0.64
Other Loc.	0.65	0.38	0.38	0.59	0.58	0.55	0.49
Score/.../Display	0.9	0.44	0.44	0.87	0.78	0.84	0.71
Small Loc.	0.55	0.29	0.29	0.5	0.46	0.47	0.34

(0.61 and 0.59), and a ratio of 0.8 obtains the worst accuracy and F1 scores (0.53 and 0.5).

Meanwhile, Table 4 and Table 5 show the results for each category averaged over games. It is clear that MST-DIM still performs the best category-wise in partially observable environments, and achieves scores that is slightly worse than ST-DIM that is probed under fully observable environments. A different masking ratio in pretraining still enhances the model’s capability across games in probing tasks, and even a masking ratio (0.8) that is remote from the probing masking ratio (0.4) can facilitate achieving better scores for each ground truth category.

6 Discussion

It was found that there was a sizable gap between the performances of ST-DIM and supervised training by [1]. However, their difference is trivial under our partially observable setting. Meanwhile, MST-DIM has demonstrated better performance than both of them. The reason might be that ST-DIM and supervised methods do not possess better initializations, and yield similarly deteriorated results in partially observable environments.

On the other hand, Randomly initialized CNNs can perform reasonably well in probing tasks. Their scores are only slightly lower than those of generative methods in [1], because random CNNs are considered a strong prior in Atari games and can capture the inductive bias [2, 25].

Different masking ratios in pretraining have shown to be effective, even for a large masking ratio that generates visually invisible images. Although the closer the pretraining masking ratio is, the more accurate the probing prediction can be, the masking ratio in pretraining is not required to be the same. The results

indicate that unsupervised pretraining can learn reliable latent generative factors from a different data distribution. Unlike in fully observable environments, this is strong evidence that contrastive methods can play a key role in strengthening model capabilities in partially observable environments for downstream tasks.

The results of small object localization in Tables 4 and 5 are highlighted in ST-DIM because generative methods typically do not penalize enough for not modeling the pixels making up small objects. The local-local contrastive task in ST-DIM is specialized in capturing local representation. In our experiment, it is clear that the performances of ST-DIM and the supervised method have dropped significantly in partially observable experiments because small objects with few pixels can be easily masked out completely. On the other hand, MST-DIM overcomes this problem by masked pretraining and is close to achieving the same level of performance (0.47 to 0.53 for F1 scores and 0.5 to 0.55 in accuracy scores). However, MST-DIM has suffered from pretraining with a masking ratio of 0.8, which masks most of the image.

In some games, such as boxing, easy-to-learn features might saturate the objective and let contrastive methods fail. For example, contrastive methods other than ST-DIM fail to model features besides the clock in boxing [1]. This is also a problem in partially observable environments and causes ST-DIM, supervised method, and MST-DIM with a masking ratio of 0.8 to perform worse. On the other hand, MST-DIM with a masking ratio close to 0.4 exhibits robustness as ST-DIM in fully observable environments.

For the study of different masking ratios, we observe that a masking ratio that is slightly higher than one of the underlying observations obtains the best accuracy and F1 scores when the original ratio is not available. Thus, it suggests that a higher masking ratio facilitates unsupervised representation learning in partially observable environments. If we were to choose between decreasing or increasing the ratio with the same amount in pretraining, increasing the number could be a reasonable choice.

7 Conclusion

We propose MST-DIM in this paper to deal with partially observable environments through pretraining. MST-DIM is a contrastive method based on an estimate of mutual information bound and uses masking in unsupervised pretraining to ensure the agent can learn reliable latent generative factors. Experiments are conducted using a benchmark of unsupervised learning on the annotated interface of Atari 2600 games. MST-DIM shows the benefit of using unsupervised representation learning in partially observable environments by achieving higher accuracy and F1 scores than ST-DIM and supervised learning.

For future work, it would be interesting to directly apply MST-DIM to RL environments and evaluate its performance using RL baselines. Designing an auxiliary contrastive loss in RL is typical, but the implementation details can vary among different research. Exploiting the weight initialization of pretrained representation encoders that resembles more to CV probing tasks can also be an intriguing topic.

Recognizing small objects can be a challenging task in partially observable environments. ST-DIM deploys the local-local contrastive task to reliably learn the representations of small objects. However, it still suffers from information loss and obtains worse results in partially observable environments. MST-DIM deals with this issue and achieves scores close to ST-DIM under the fully observable setup. It still remains to be studied how to recognize small objects where the environment is almost invisible, and information loss is severe.

Acknowledgements. This work was performed on the [ML node] resource, owned by the University of Oslo, and operated by the Department for Research Computing at USIT, the University of Oslo IT-department. <http://www.hpc.uio.no/>.

References

1. Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.A., Hjelm, R.D.: Unsupervised state representation learning in atari. In: *Advances in Neural Information Processing Systems*, vol. 32 (2019)
2. Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., Efros, A.A.: Large-scale study of curiosity-driven learning. arXiv preprint [arXiv:1808.04355](https://arxiv.org/abs/1808.04355) (2018)
3. Cassandra, A.R., Kaelbling, L.P., Littman, M.L.: Acting optimally in partially observable stochastic domains. In: *AAAI*, vol. 94, pp. 1023–1028 (1994)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, pp. 1597–1607. PMLR (2020)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
7. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060 (2017)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
9. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: DRAW: a recurrent neural network for image generation. In: *International Conference on Machine Learning*, pp. 1462–1471. PMLR (2015)
10. Guo, Z.D., et al.: Bootstrap latent-predictive representations for multitask reinforcement learning. In: *International Conference on Machine Learning*, pp. 3875–3886. PMLR (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738 (2020)
12. Hjelm, R.D., et al.: Learning deep representations by mutual information estimation and maximization. arXiv preprint [arXiv:1808.06670](https://arxiv.org/abs/1808.06670) (2018)
13. Jonschkowski, R., Brock, O.: Learning state representations with robotic priors. *Auton. Robot.* **39**(3), 407–428 (2015). <https://doi.org/10.1007/s10514-015-9459-7>

14. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114) (2013)
15. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1920–1929 (2019)
16. Laskin, M., Srinivas, A., Abbeel, P.: CURL: contrastive unsupervised representations for reinforcement learning. In: International Conference on Machine Learning, pp. 5639–5650. PMLR (2020)
17. Lee, K.H., et al.: Predictive information accelerates learning in RL. *Adv. Neural. Inf. Process. Syst.* **33**, 11890–11901 (2020)
18. Lesort, T., Díaz-Rodríguez, N., Goudou, J.F., Filliat, D.: State representation learning for control: an overview. *Neural Netw.* **108**, 379–392 (2018)
19. Mnih, V., et al.: Playing atari with deep reinforcement learning (2013)
20. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748) (2018)
21. Puterman, M.L.: Markov decision processes. *Handbooks Oper. Res. Management Sci.* **2**, 331–434 (1990)
22. Stooke, A., Lee, K., Abbeel, P., Laskin, M.: Decoupling representation learning from reinforcement learning. In: International Conference on Machine Learning, pp. 9870–9879. PMLR (2021)
23. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge (2018)
24. Taylor, L.N., Whalen, Z.: *Playing the Past: History and Nostalgia in Video Games*. JSTOR (2008)
25. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2018)
26. Watkins, C.J., Dayan, P.: Q-learning. *Mach. Learn.* **8**(3–4), 279–292 (1992)
27. Zhu, J., et al.: Masked contrastive representation learning for reinforcement learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 3421–3433 (2022)