A Simple Framework for Self-Supervised Learning of Sample-Efficient World Models

Jan Robine[*], **Marc Höftmann, Stefan Harmeling**
Department of Computer Science
Technical University of Dortmund

Abstract

Deep reinforcement learning algorithms suffer from low sample efficiency, which is addressed in recent approaches by building a world model and learning behaviors in imagination. We present a simple framework for self-supervised learning of world models inspired by VICReg, requiring neither image reconstructions nor specific neural network architectures. The learned representations are temporally consistent, which facilitates next state prediction and leads to good generalization properties for the policy and the value function. We build a world model for Atari consisting only of feedforward layers that is easy to implement and allows fast training and inference. By learning behaviors in imagination, we evaluate our method on the Atari 100k benchmark.

1 Introduction

Deep reinforcement learning has shown great success on challenging decision making problems [32, 39, 33, 22, 3, 37, 26, 18]. However, sample efficiency remains the biggest challenge for reinforcement learning algorithms, i.e., the amount of data that is required to learn good behaviors. Recent works increase the sample efficiency with improved architectural design and hyperparameters [46, 42], borrowing ideas from representation learning [12, 29, 40, 41, 42], data augmentation [49, 28], pretraining and fine-tuning [41], or with learned a model of the environment [25, 50, 34, 36, 31, 16, 17, 18].

Dyna [43] introduced the idea of learning a model of the environment to improve the value function. Ha and Schmidhuber [15] learn a *world model*, which is a deep generative model of the environment's dynamics and rewards. By imagining trajectories in a compact space, improved behaviors can be learned without further environment interactions. The representations are obtained by a variational autoencoder [27] and the dynamics are modeled with an LSTM [23]. Hafner et al. [16, 17, 18] jointly train a variational autoencoder and a recurrent neural network. They achieve good performance across multiple domains with discrete state representations and carefully chosen objective functions. Micheli et al. [31], Robine et al. [36] model the environment with transformers [47] and achieve state-of-the-art results on the Atari 100k benchmark.

In computer vision, self-supervised learning of image representations has made significant progress in recent years [8, 20, 14, 6, 9, 51, 7, 4]. Many approaches are based on a Siamese architecture [5] and can be divided into contrastive and non-contrastive methods. Contrastive methods [8, 20] learn representations that are similar for different views (e.g. image augmentations) of the same image (positive), but dissimilar for different images (negative) to prevent a collapse of the representations. Non-contrastive methods do not rely on negative images, but rather prevent representation collapse by the design of the architecture [14, 9] or by regularization of the representations [51, 4].

[*]Correspondence to jan.robine@tu-dortmund.de

37th Conference on Neural Information Processing Systems (NeurIPS 2023), Workshop on Self-Supervised Learning.

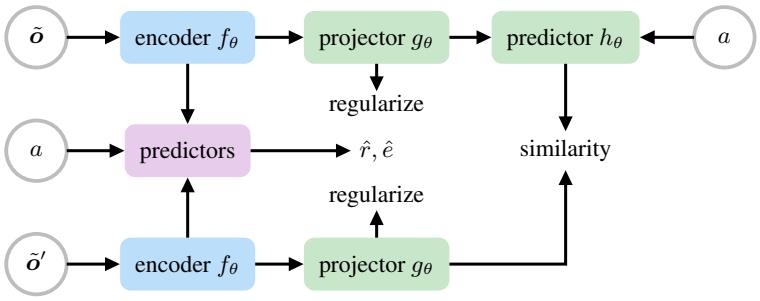

Figure 1: The representations of our world model are learned using a self-supervised framework inspired by VICReg [4].

In this work, we introduce a simple framework for learning world models for imagination based on recent advances in self-supervised learning. The contributions of our work are as follows:

- We implement a world model inspired by the self-supervised VICReg framework [4] without the need for image reconstruction and evaluate it on the Atari 100k benchmark [25]. To the best of our knowledge, learning by imagination without image reconstruction has not been (successfully) applied to Atari games [16, 34]. Note, however, that lookahead search algorithms have already been successful without reconstructions [37, 50], and that model-based RL without reconstructions in general has already been studied [35, 12, 34, 19, 13].
- Our learned representations are temporally consistent and continuous. The representations of previous world models used for simulation on Atari games were categorical [17, 18, 31, 36]. The continuity has several advantages: (i) good generalization properties for the policy and value function, (ii) stable training since jumps in the distribution of the representations are less likely, (iii) no need for straight-through gradient estimation.
- Our world model does not rely on a specific architectural design and is easy to implement; we only use feedforward layers. Nonetheless, it can be easily extended to more complex architectures, e.g., a recurrent or stochastic dynamics predictor.

2 Method

2.1 World Model

We formalize the environment in terms of a partially observable Markov decision process (POMDP) with discrete time steps, rewards $r \in \mathbb{R}$, image observations $\boldsymbol{o} \in \mathbb{R}^{C \times H \times W}$, and discrete actions $a \in \mathbb{N}$. A transition inside the environment is described by a tuple $(\boldsymbol{o}, \boldsymbol{a}, r, e, \boldsymbol{o}')$, where $e \in \{0, 1\}$ indicates episode ends.

The task of our world model is threefold: (i) Map image observations onto compact representations that retain relevant features, (ii) predict rewards and episode ends, and (iii) predict next states in representation space. These three components are used to synthesize trajectories efficiently.

We employ a self-supervised representation learning approach inspired by VICReg [4] to learn an encoder f_θ that extracts temporally consistent features from the observations and maximizes their information content; we give an overview in Figure 1. Given a transition $(\boldsymbol{o}, \boldsymbol{a}, r, \boldsymbol{o}')$ of the POMDP, we apply random image augmentations $t, t' \sim \mathcal{T}$, sampled from a predefined set $\mathcal{T}$, to obtain augmented observations $\tilde{\boldsymbol{o}} = t(\boldsymbol{o})$ and $\tilde{\boldsymbol{o}}' = t'(\boldsymbol{o}')$. A Siamese encoder f_θ computes representations $\tilde{\boldsymbol{y}} = f_\theta(\tilde{\boldsymbol{o}})$ and $\tilde{\boldsymbol{y}}' = f_\theta(\tilde{\boldsymbol{o}}')$ with $\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{y}}' \in \mathbb{R}^d$. A Siamese projector g_θ computes embeddings $\tilde{\boldsymbol{z}} = g_\theta(\tilde{\boldsymbol{y}})$ and $\tilde{\boldsymbol{z}}' = g_\theta(\tilde{\boldsymbol{y}}')$ with $\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}' \in \mathbb{R}^D$. An embedding predictor h_θ predicts the next embedding $\hat{\boldsymbol{z}}' = h_\theta(\tilde{\boldsymbol{z}}, a)$. To achieve temporal consistency between representations, we maximize the similarity between $\hat{\boldsymbol{z}}'$ and $\tilde{\boldsymbol{z}}'$ by minimizing the mean squared error. To prevent representation collapse, the embeddings are regularized using the variance and covariance regularization terms proposed by Bardes et al. [4]. We train a reward predictor $p_\theta(r \mid \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{y}')$ via discrete regression using two-hot encoded targets and symlog predictions, as proposed by Hafner et al. [18]. Thus, reward prediction is stable across different scales without the need for normalization. We train an episode end predictor $p_\theta(e \mid \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{y}')$ via binary classification. Note, that the reward predictor and episode end predictor are conditioned on the next representation $\boldsymbol{y}'$, which facilitates the prediction tasks.

The rewards and episode ends provide stable supervised training signals, so we jointly minimize the losses of the encoder and the predictors, which leads to the final loss

$$\mathcal{L}(\theta) = \mathbb{E}_\tau \Big[\underbrace{\lambda \tfrac{1}{D} \|\hat{\mathbf{z}}' - \tilde{\mathbf{z}}'\|_2^2}_{\text{Similarity}} + \underbrace{\mathrm{VC}(\tilde{\mathbf{Z}}) + \mathrm{VC}(\tilde{\mathbf{Z}}')}_{\text{Regularization}} - \underbrace{\log p_\theta(\mathrm{r} \,|\, \tilde{\mathbf{y}}, \mathrm{a}, \tilde{\mathbf{y}}')}_{\text{Reward predictor}} - \underbrace{\log p_\theta(\mathrm{e} \,|\, \tilde{\mathbf{y}}, \mathrm{a}, \tilde{\mathbf{y}}')}_{\text{Episode end predictor}} \Big], \quad (1)$$

where τ is a batch of transitions from a replay buffer, $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{Z}}'$ are batches of embeddings, $\lambda > 0$ controls the strength of the similarity loss, and VC is VICReg's variance-covariance loss [4], i.e.,

$$\mathrm{VC}(\mathbf{Z}) = \frac{1}{D} \sum_{j=1}^{D} \Bigg[\underbrace{\mu \max\Big(0, 1 - \sqrt{\mathrm{Cov}(\mathbf{Z})_{j,j} + \epsilon}\Big)}_{\text{Variance regularization}} + \underbrace{\nu \sum\nolimits_{k \neq j} \mathrm{Cov}(\mathbf{Z})_{j,k}^2}_{\text{Covariance regularization}} \Bigg], \quad (2)$$

where D is the dimensionality of the embeddings, $\mu, \nu > 0$ control the strength of variance and covariance regularization terms, respectively, and $\epsilon > 0$ prevents numerical instabilities.

Being detached from representation learning, a dynamics predictor q_ϕ learns to predict the next representation $\hat{\boldsymbol{y}}' = q_\phi(\boldsymbol{y}, a)$ by minimizing the mean squared error

$$\mathcal{L}(\phi) = \mathbb{E}_\tau \Big[\tfrac{1}{d} \|\hat{\mathbf{y}}' - \mathbf{y}\|_2^2 \Big]. \quad (3)$$

Note, that we train the dynamics predictor with non-augmented observations, i.e., $\boldsymbol{y} = f_\theta(\boldsymbol{o})$ and $\boldsymbol{y}' = f_\theta(\boldsymbol{o}')$ instead of $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}'$, to avoid the noise introduced by the augmentations, which are only necessary for representation learning. Furthermore, we exploit the temporal consistency of the representations and add a skip connection to the dynamics predictor q_ϕ, so that it only has to learn the change of representations $\mathbb{E}[\mathbf{y}' - \mathbf{y} \,|\, \boldsymbol{y}, \boldsymbol{a}]$. We hypothesize that the distribution of $\mathbf{y}$ changes more rapidly during training than the difference between $\mathbf{y}$ and $\mathbf{y}'$, since the objective of the representation model is to keep the representations close to each other but not close to some prior distribution.

2.2 Policy

The representations $\boldsymbol{y}$ serve as states for the policy $\pi_\psi(\boldsymbol{a} \,|\, \boldsymbol{y})$. The policy learns to maximize the expected return for each state by performing approximate gradient ascent with the policy gradient [45]. We reduce the variance of the gradient estimates with a learned value function $v_\xi(\boldsymbol{y})$ as baseline, resulting in an advantage actor-critic approach [33]. The world model simulates batches of sequences of length $h = 10$, which are used to estimate the advantages with generalized advantage estimation [38] and to calculate multi-step truncated λ-returns [44] as target for the value function. Instead of advantage normalization, we adopt the strategy of Hafner et al. [18] and normalize the returns for advantage computation by mapping the 5$^{\text{th}}$ and the 95$^{\text{th}}$ percentile to 0 and 1, respectively. We add the entropy of the policy to the objective to improve exploration, as it prevents early convergence to suboptimal policies [48, 33]. For the value function we use the same discrete regression approach as for the reward predictor, i.e., two-hot encoded targets and symlog predictions [18].

2.3 Implementation Details

The Siamese encoder is implemented by an ImpalaCNN [11], which outputs representations of dimension $d = 512$. The Siamese projector and the embedding predictor are MLPs with hidden dimensions $2048{-}2048$, computing embeddings of dimension $D = 2048$. The reward predictor, episode end predictor, policy, and value function are MLPs with hidden dimensions $1024{-}1024$. All networks use SiLU nonlinearities [21], the encoder uses batch normalization [24], and the MLPs use layer normalization [2]. We use the AdamW optimizer [30] for all networks and loss functions.

The representations are continuous and not regularized directly—unlike variational autoencoders [27], for example, which force the representations to be close to a standard normal distribution. To stabilize training, Schrittwieser et al. [37] and Schwarzer et al. [40] normalize the representations to lie in the interval $[0, 1]$. We found that adding another normalization layer at the end of the encoder is sufficient and ensures that the norm of the representations stays roughly constant during training.

We use the image augmentations proposed by Yarats et al. [49], i.e., random shifts and imagewise intensity jittering. We stack the four most recent frames [32], so that the world model can encode the velocity of objects into the representations. Moreover, we stack the four most recent actions, which are fed into the predictors. This is especially helpful for our feedforward world model, since the effects of actions might be slightly delayed.

Table 1: Comparison with other world models used for imagination.

Setting	**SimPLe** [25]	**IRIS** [31]	**TWM** [36]	**DreamerV3** [18]	**Ours**
Discrete Representation	x	x	x	x	
Decoder	x	x	x	x	
Pixel Imagination	x	x			
Pixel Dynamics	x				
Sequential Dynamics		x	x	x	
Data Augmentation					x

Table 2: Comparison with other methods on the Atari100k benchmark.

			Model-free	Lookahead	Imagination		
Game	**Random**	**Human**	**SPR**	**Eff. Zero**	**IRIS**	**DreamerV3**	**Ours**
Alien	227.8	7127.7	841.9	808.5	420.0	959	505.1
Amidar	5.8	1719.5	179.7	148.6	143.0	139	84.2
Assault	222.4	742.0	565.6	1263.1	1524.4	706	537.5
Asterix	210.0	8503.3	962.5	25557.8	853.6	932	1054.8
Bank Heist	14.2	753.1	345.4	351.0	53.1	649	22.3
Battle Zone	2360.0	37187.5	14834.1	13871.2	13074.0	12250	4800.0
Boxing	0.1	12.1	35.7	52.7	70.1	78	90.5
Breakout	1.7	30.5	19.6	414.1	83.7	31	40.3
Chopper Cmd.	811.0	7387.8	946.3	1117.3	1565.0	420	1967.8
Crazy Climber	10780.5	35829.4	36700.5	83940.2	59324.2	97190	25353.2
Demon Attack	152.1	1971.0	517.6	13003.9	2034.4	303	1107.7
Freeway	0.0	29.6	19.3	21.8	31.1	0	17.7
Frostbite	65.2	4334.7	1170.7	296.3	259.1	909	365.8
Gopher	257.6	2412.5	660.6	3260.3	2236.1	3730	2515.9
Hero	1027.0	30826.4	5858.6	9315.9	7037.4	11161	2536.0
James Bond	29.0	302.8	366.5	517.0	462.7	445	289.3
Kangaroo	52.0	3035.0	3617.4	724.1	838.2	4098	1465.2
Krull	1598.0	2665.5	3681.6	5663.3	6616.4	7782	6432.74
Kung Fu Master	258.5	22736.3	14783.2	30944.8	21759.8	21420	12464.0
Ms Pacman	307.3	6951.6	1318.4	1281.2	999.1	1327	1217.5
Pong	-20.7	14.6	-5.4	20.1	14.6	18	11.6
Private Eye	24.9	69571.3	86.0	96.7	100.0	882	61.6
Qbert	163.9	13455.0	866.3	13781.9	745.7	3405	688.5
Road Runner	11.5	7845.0	12213.1	17751.3	9614.6	15565	6542.4
Seaquest	68.4	42054.7	558.1	1100.2	661.3	618	344.2
Up N Down	533.4	11693.2	10859.2	17264.2	3546.2	NaN	3165.0
Normalized Mean	0.000	1.000	0.616	1.943	1.046	1.12	0.826
Normalized Median	0.000	1.000	0.396	1.090	0.289	0.49	0.355

3 Experiments and Discussion

We evaluate our world model on the Atari 100k benchmark, which was first proposed by Kaiser et al. [25] and has been used to evaluate many sample-efficient reinforcement learning methods [29, 49, 40, 31, 18]. It includes a subset of 26 Atari games and is limited to 400k environment steps, which amounts to 100k steps after frame skipping or roughly 2 hours of gameplay. We perform 5 runs per game and for each run we compute the average score over 100 episodes at the end of training. In Table 2 we compare our method with five baselines: the model-free algorithm SPR [40] (scores from Agarwal et al. [1]) and the model-based methods EfficientZero [50], IRIS [31], and DreamerV3 [18]. The aggregate metrics are computed on human normalized scores.

We presented a simple approach for self-supervised learning of world models, with significant differences to previous methods, as summarized in Table 1. We successfully apply our world model to the Atari 100k benchmark. The application to other environments is left for future work. Since VICReg [4] is capable to learn state-of-the-art representations on the visually complex ImageNet dataset [10], we suppose that our proposed framework should generally work in more complex environments, where image reconstruction is difficult.

References

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29304–29320, 2021.

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[3] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 507–517. PMLR, 2020.

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 737–744. Morgan Kaufmann, 1993.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE, 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[11] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1406–1415. PMLR, 2018.

[12] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179. PMLR, 2019.

[13] Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Russ Salakhutdinov. Simplifying model-based RL: learning representations, latent-space models, and policies with one objective. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[15] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2455–2467, 2018.

[16] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[17] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[18] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse domains through world models. *CoRR*, abs/2301.04104, 2023.

[19] Nicklas Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 8387–8406. PMLR, 2022.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

[21] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016.

[22] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3215–3222. AAAI Press, 2018.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, 1997.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.

[25] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[26] Steven Kapturowski, Victor Campos, Ray Jiang, Nemanja Rakicevic, Hado van Hasselt, Charles Blundell, and Adrià Puigdomènech Badia. Human-level atari 200x faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[28] Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[29] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 2020.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[31] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015.

[33] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1928–1937. JMLR.org, 2016.

[34] Tung D. Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8130–8139. PMLR, 2021.

[35] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep RL for model-based control. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[36] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[37] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nat.*, 588(7839):604–609, 2020.

[38] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

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

[40] Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron C. Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[41] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. *CoRR*, abs/2106.04799, 2021.

[42] Max Schwarzer, Johan Samir Obando-Ceron, Aaron C. Courville, Marc G. Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level atari with human-level efficiency. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 30365–30380. PMLR, 2023.

[43] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bull.*, 2(4):160–163, 1991.

[44] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

[45] Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063. The MIT Press, 1999.

[46] Hado van Hasselt, Matteo Hessel, and John Aslanides. When to use parametric models in reinforcement learning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14322–14333, 2019.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[48] Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

[49] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[50] Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games with limited data. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 25476–25488, 2021.

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021.