DeepMind

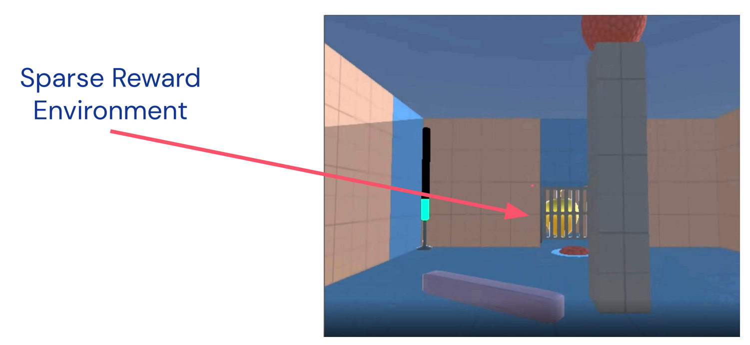
BYOL-Explore: Exploration by Bootstrapped Prediction

Zhaohan Daniel Guo*, Shantanu Thakoor*, Miruna Pislar*, Bernardo Avila Pires*, Florent Altché*, Corentin Tallec*, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, Michal Valko, Remi Munos, Mohammad Gheshlaghi Azar*, Bilal Piot*

*Equal Contribution

Motivation

Exploration is hard in large, visually complex domains



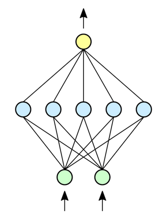
Too many states to try to explore everything!

THEREFORE

We must focus on only exploring the **interesting** parts

Curiosity-Driven Exploration

Build a world model



Explore the mistakes of world model to better refine it

$$\max_{\pi} \ \text{WorldModelLoss}(\pi)$$

The world model determines what is interesting to explore and what to ignore

Our Contribution: BYOL-Explore

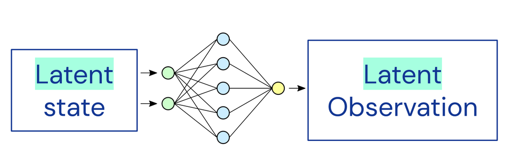
We believe

Strong Representations —Lead to→ Strong Exploration

Strong Exploration —Leads to→ Strong Representations

Approach:

Extend BYOL[1] to learn a latent dynamics model



Explore the latent mistakes of the world model to better refine it

$$\underbrace{\max_{\pi}}_{\text{Exploration}} \ \underbrace{\min_{\theta}}_{\text{Dynamics Model}} \ \text{BYOLLoss}_{\pi}(\theta)$$

Exploration Dynamics Model Learning

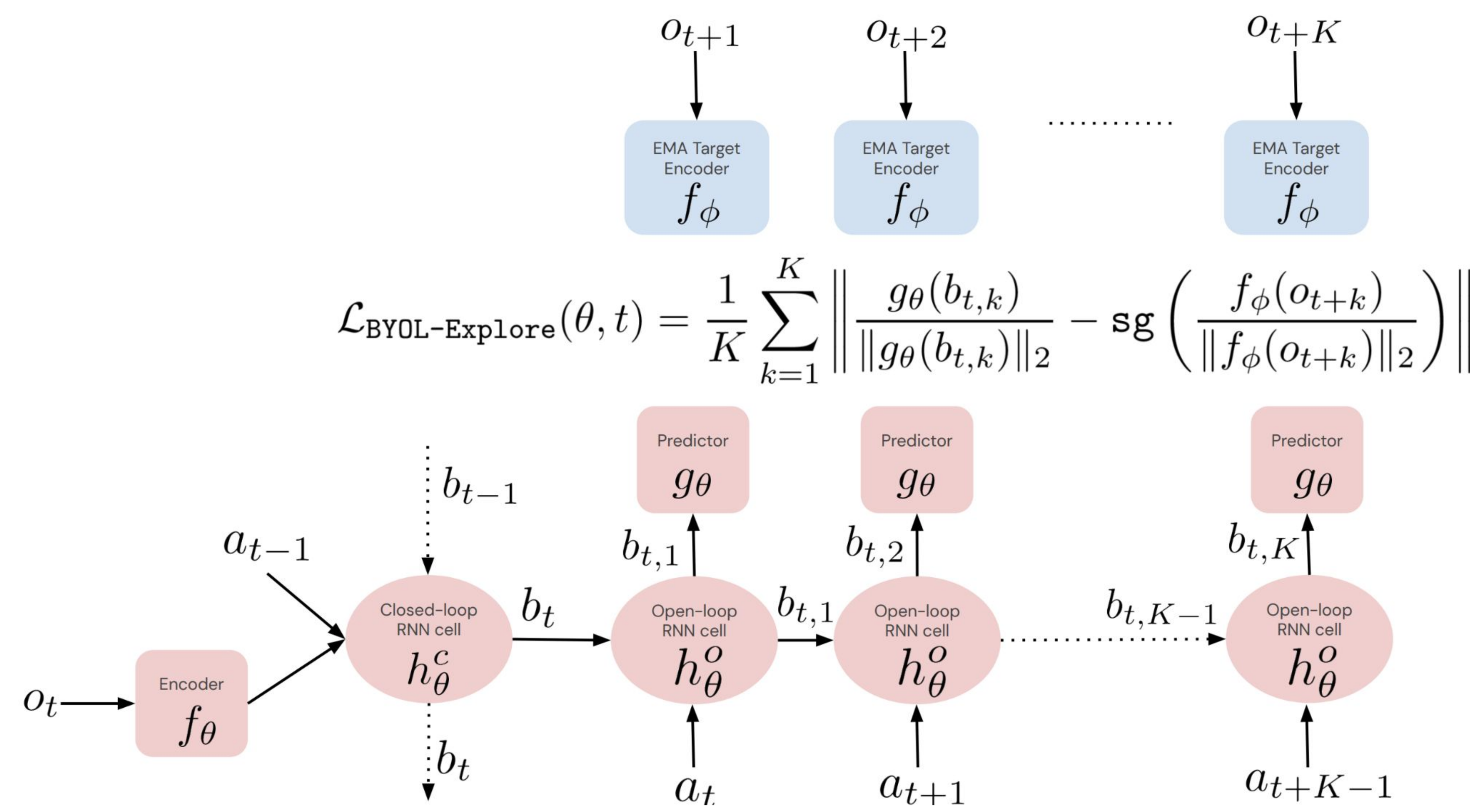
The mistakes are dynamics-aware and structured, since they are in latent space

One unified objective for representation learning, dynamics modelling, and exploration

[1] Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, Piot B. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems. 2020;33:21271-84.

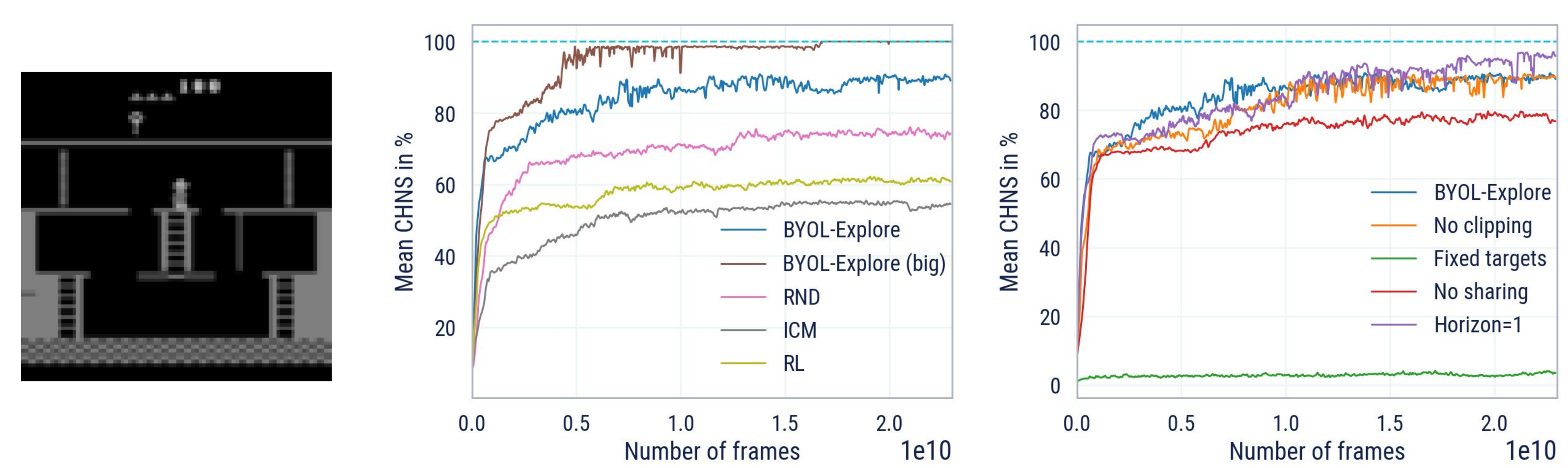
BYOL-Explore Algorithm

1. Encode observations o_t into latents with f_θ
2. Compress the history of observations and actions into b_t with a closed-loop RNN (h^c_θ)
3. (**World Model**) Combine b_t, and future actions with an open-loop RNN (h^o_θ) and pass through a predictor g_θ to predict the corresponding future latent observation
4. (**Prediction Targets**) Encode future observations $o_{t+1}, \ldots, o_{t+K}$ with the target network f_ϕ (EMA of f_θ)
5. (**Mistakes**) Compute the normalized L_2 (cosine similarity) loss (stopping gradients to targets)
6. (**Intrinsic Reward**) Standardize and ReLu the loss to use it as the intrinsic reward



$$\mathcal{L}_{\texttt{BYOL-Explore}}(\theta, t) = \frac{1}{K}\sum_{k=1}^{K}\left\| \frac{g_\theta(b_{t,k})}{\|g_\theta(b_{t,k})\|_2} - \text{sg}\left(\frac{f_\phi(o_{t+k})}{\|f_\phi(o_{t+k})\|_2}\right)\right\|_2^2$$

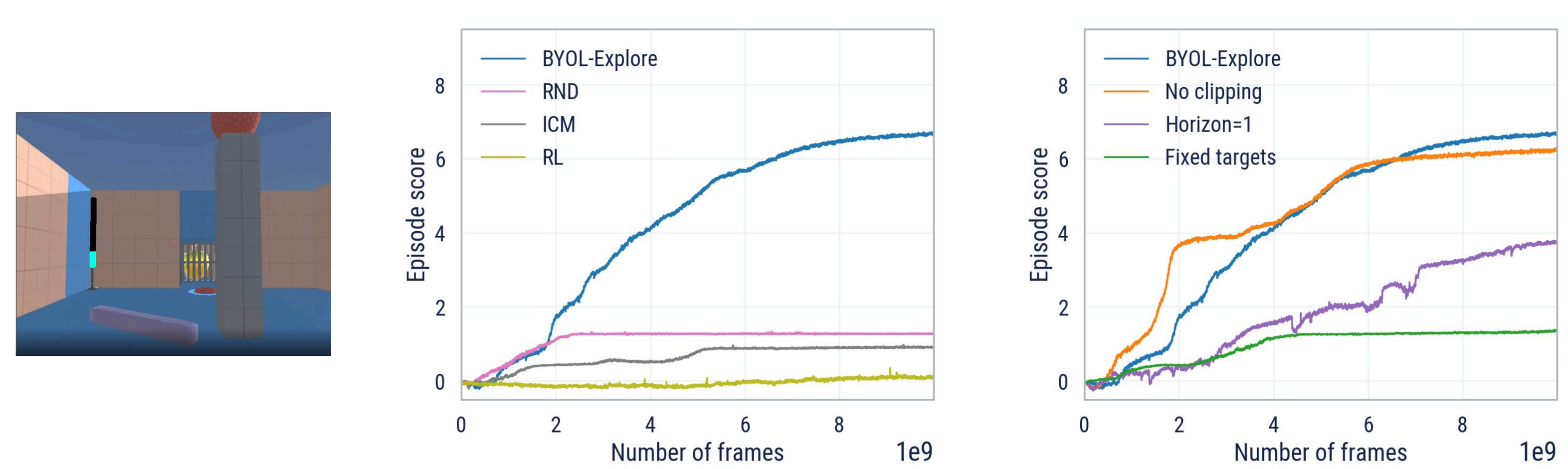
Hard Exploration Atari



Main Findings:

- BYOL-Explore greatly outperforms RND and ICM baselines in the 10 hardest exploration Atari games (in terms of clipped human-normalized score)
- Enriching the target latent representations is crucial to good performance. In contrast, predicting untrained, randomly initialized targets does not work
- Sharing the representation with RL also significantly helps performance

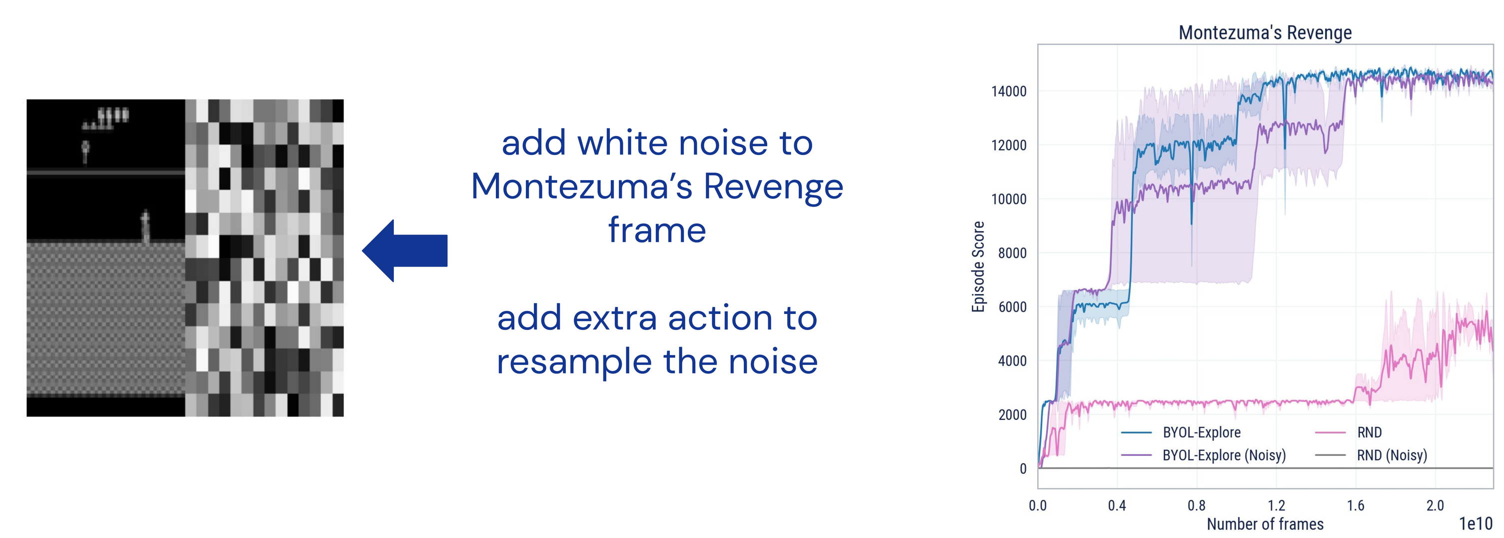
DM-Hard-8



Main Findings:

- BYOL-Explore greatly outperforms RND and ICM in multi-task DM-Hard-8, a set of partially-observable, procedurally-generated 3D navigation and puzzle tasks
- Enriching the target latent representations is crucial to good performance
- The prediction horizon is very important in a partially observable domain

Ablation: Controllable TV Noise



Main Findings:

- BYOL-Explore (purple) is completely robust to this extra controllable noise and matches the noise-free performance (blue).
- RND (pink) no longer takes off with this kind of noise.

Conclusion

- BYOL-Explore is a simple **curiosity-driven** algorithm for **jointly** doing
  - Representation learning
  - Latent Dynamics modelling
  - Exploration
- BYOL-Explore **outperforms** previous exploration methods in **diverse, visually complex domains** (Hard Exploration Atari and DM-Hard-8)
- BYOL-Explore is **robust to simple** kinds of **noise** due to operating in latent space and learning a representation that filters out the noise
- (Limitation) BYOL-Explore relies on deterministic dynamics
  - Follow-up Deep RL workshop paper that makes it robust to stochastic dynamics: *"BLaDE: Robust Exploration via Diffusion Models"*

See paper for more detailed descriptions and experimental results!