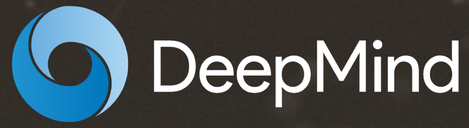
Reinforcement Learning with Unsupervised Auxiliary Tasks

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki
Tom Schaul, Joel Z Leibo, David Silver, Koray Kavukcuoglu

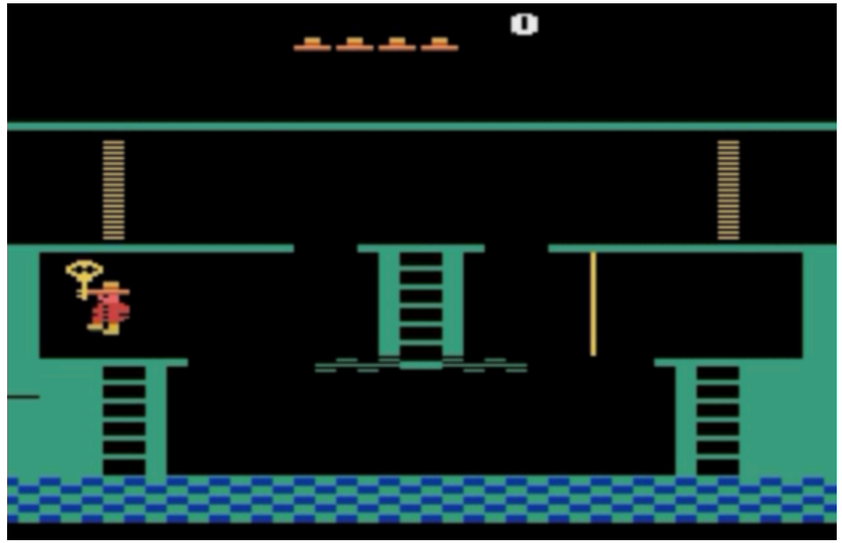


25th April 2017, ICLR

Motivation

Deep reinforcement learning is very data hungry.

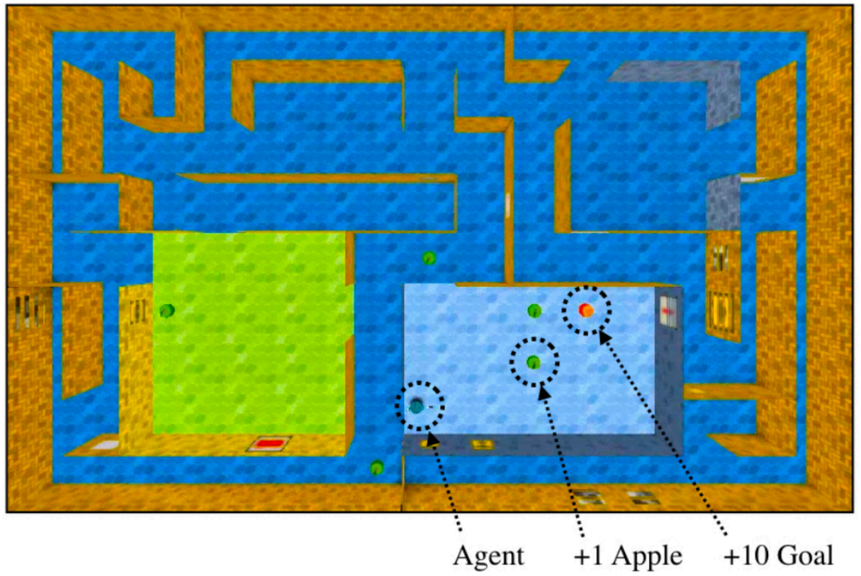
- Many steps to find sparse rewards.
- Scalar supervision (returns).
- Neural networks with many parameters.
- Slow stochastic gradient descent.



This work — **augment an RL agent with auxiliary prediction and control tasks.**

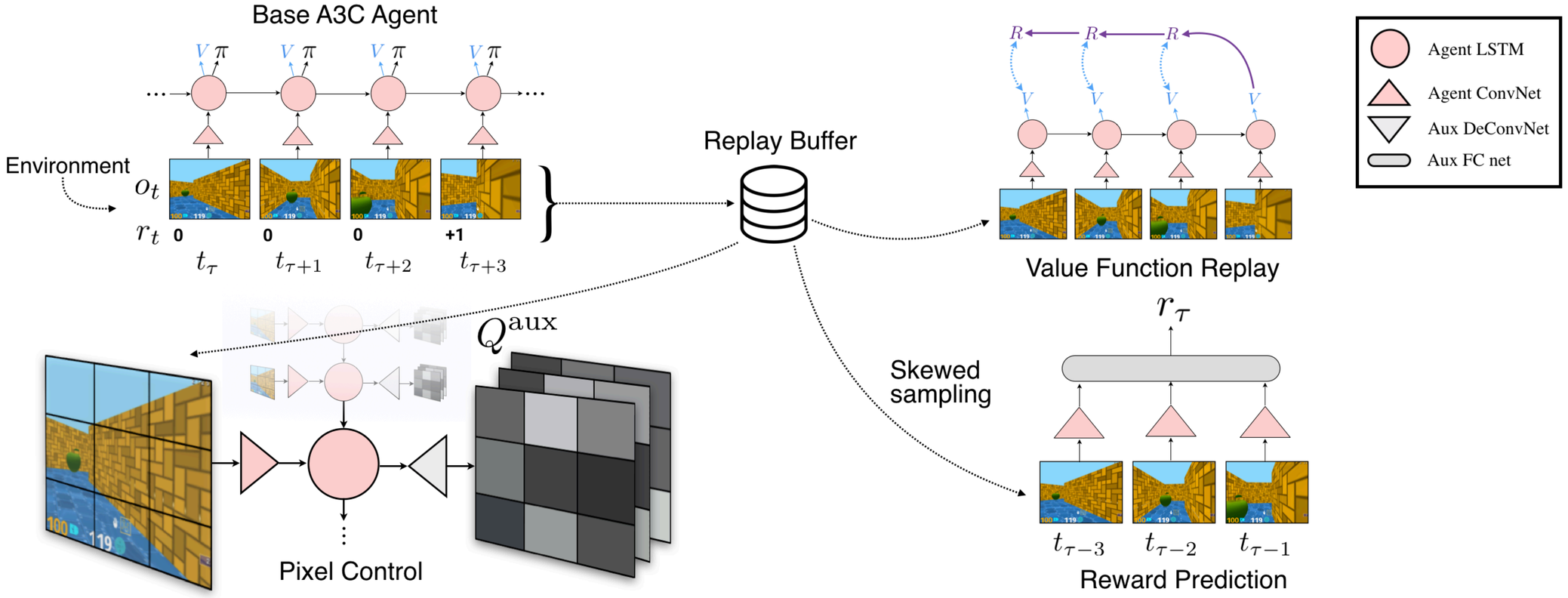
This provides extra supervision to train feature extractors resulting in

- **10x improvement in data efficiency** over A3C on 3D DeepMind Lab.
- **60% improvement in final scores** over A3C.



UNREAL

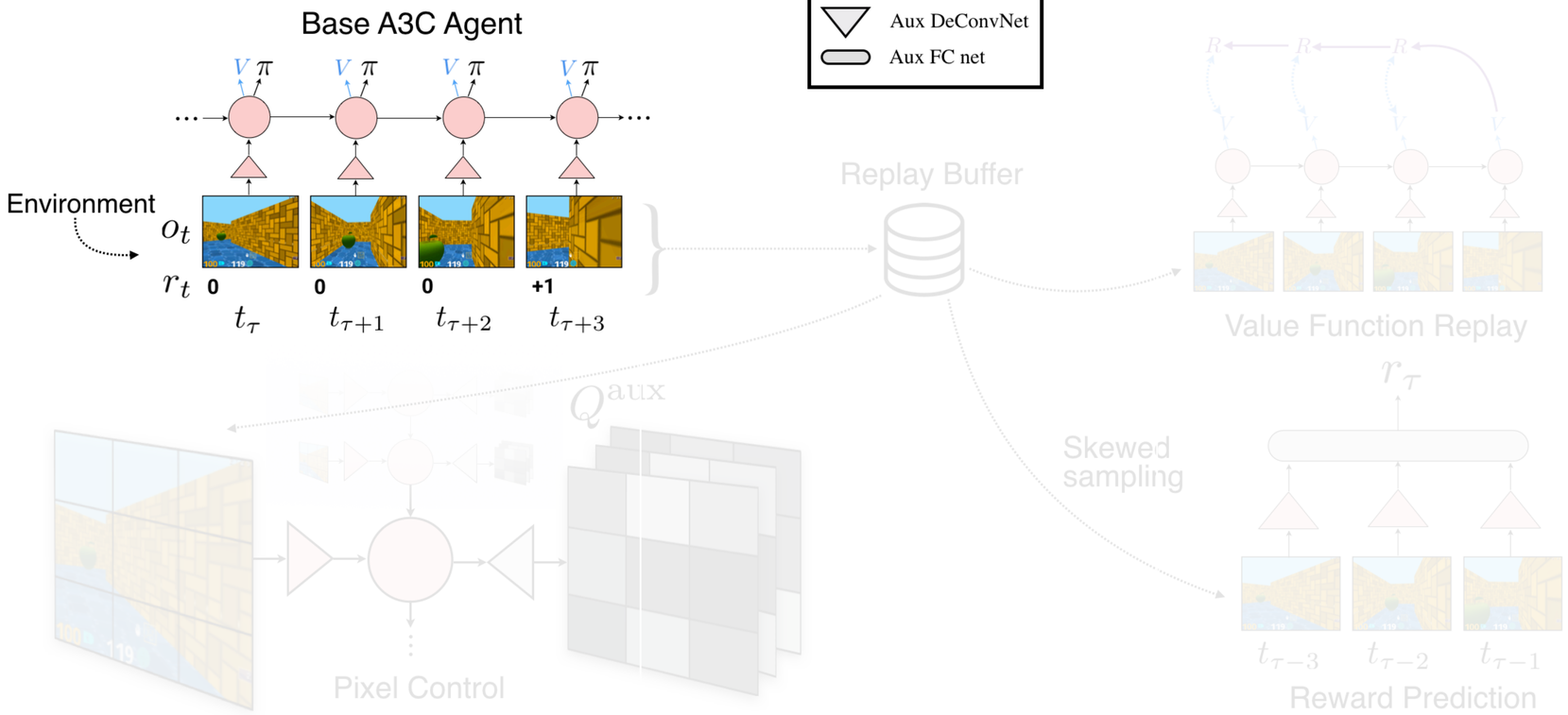
UNsupervised **RE**inforcement and **A**uxiliary **L**earning = **UNREAL** agent



UNREAL augments an LSTM A3C agent with **3 auxiliary tasks**.

Can be used on top of DQN, DDPG, TRPO, or other agents.

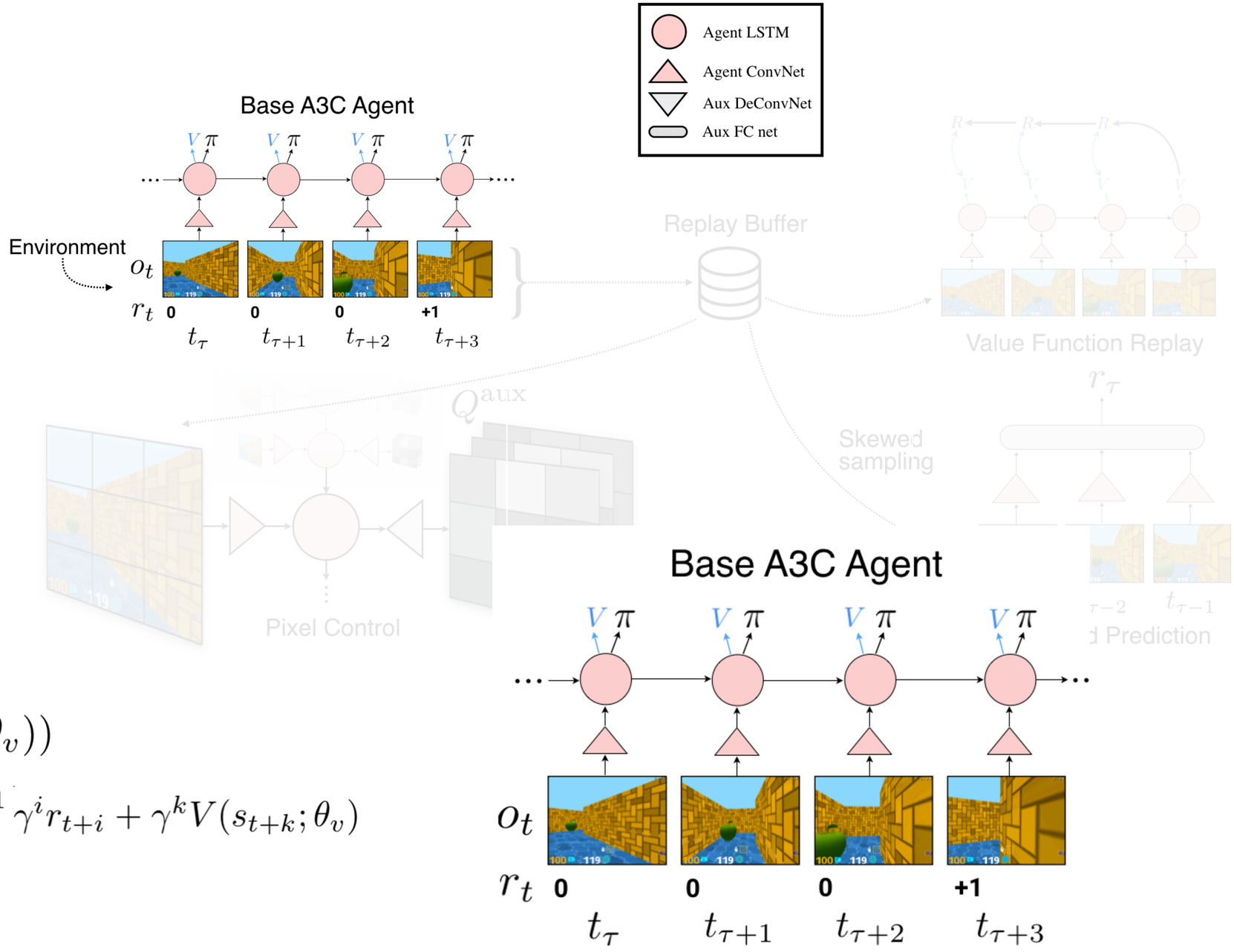
Base Policy


Agent LSTM
Agent ConvNet
Aux DeConvNet
Aux FC net
Base A3C Agent
Environment
o_t
r_t
0
0
0
+1
t_τ
$t_{\tau+1}$
$t_{\tau+2}$
$t_{\tau+3}$
Replay Buffer
Value Function Replay
Q^{aux}
Pixel Control
Skewed sampling
r_τ
$t_{\tau-3}$
$t_{\tau-2}$
$t_{\tau-1}$
Reward Prediction

Base Policy

Base policy is an LSTM agent trained with A3C *[Mnih 2016]*.

Advantage actor-critic algorithm with multiple asynchronous workers.



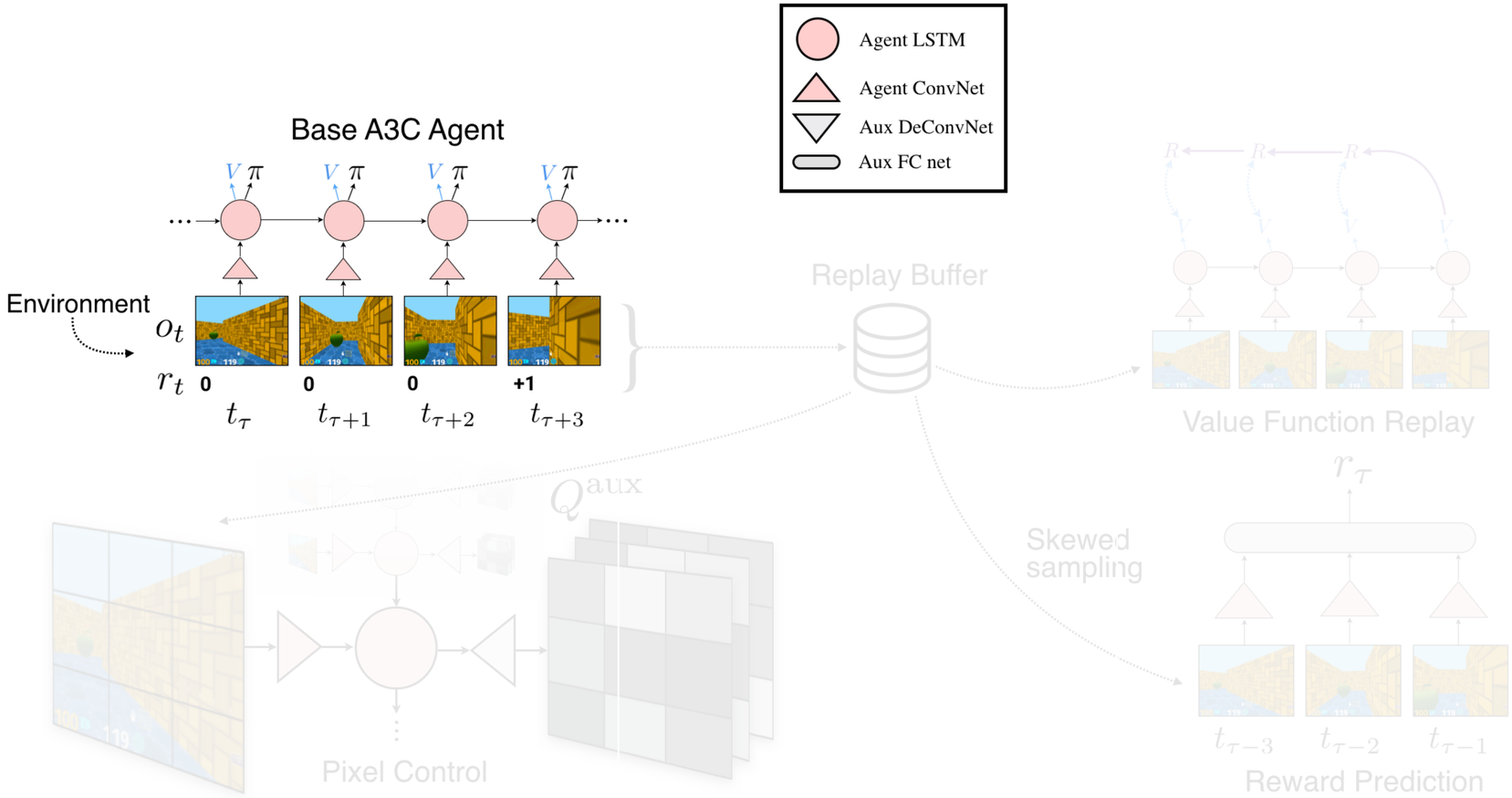
$$\nabla_\theta \log \pi(a_t|s_t;\theta)(R_t - V(s_t;\theta_v))$$

$$\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k};\theta_v)$$

Base Policy

Base policy is an LSTM agent trained with A3C *[Mnih 2016]*.

Advantage actor-critic algorithm with multiple asynchronous workers.

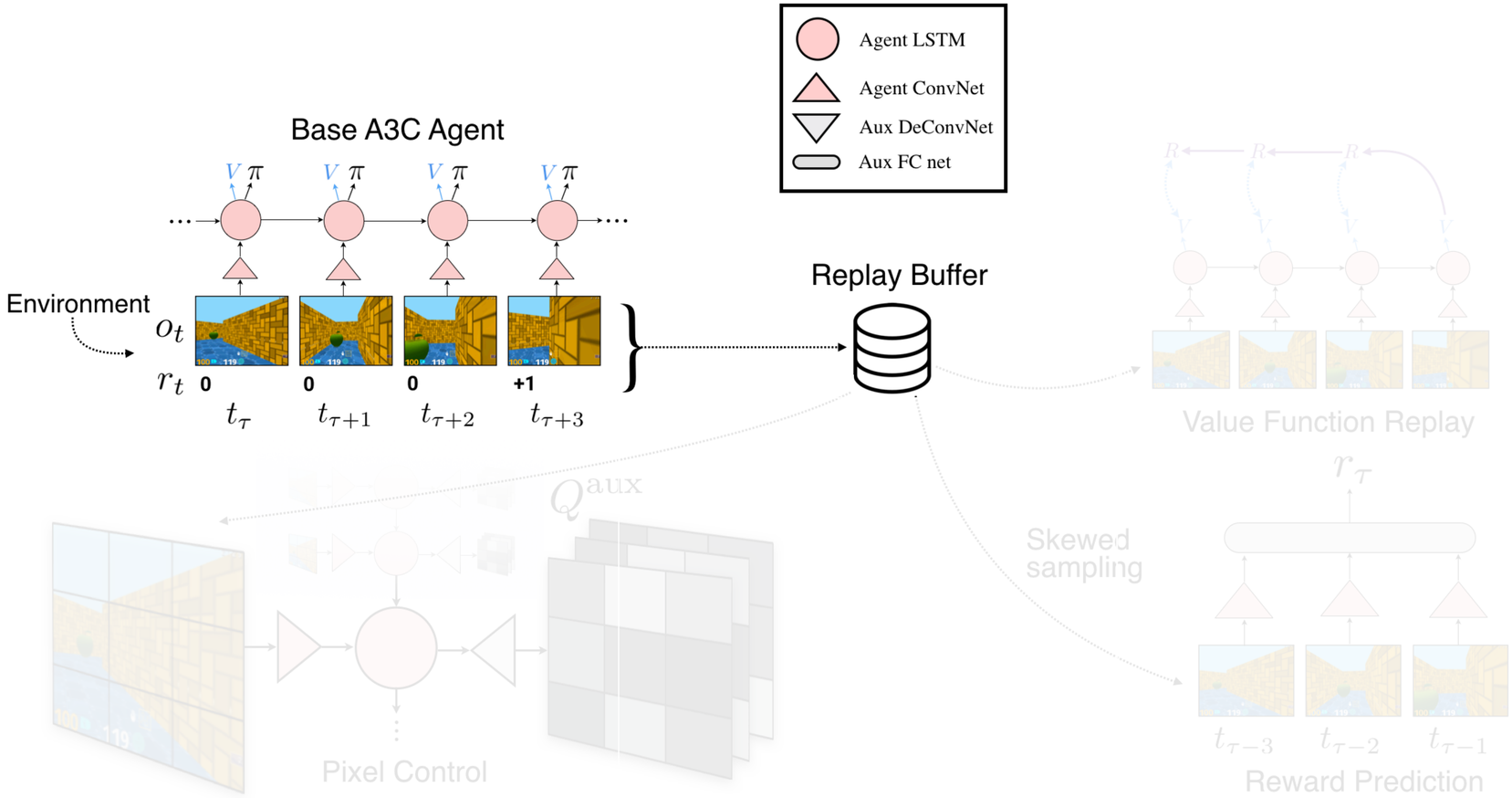


UNREAL always acts in the environment with the base A3C policy.

Base Policy

Base policy is an LSTM agent trained with A3C *[Mnih 2016]*.

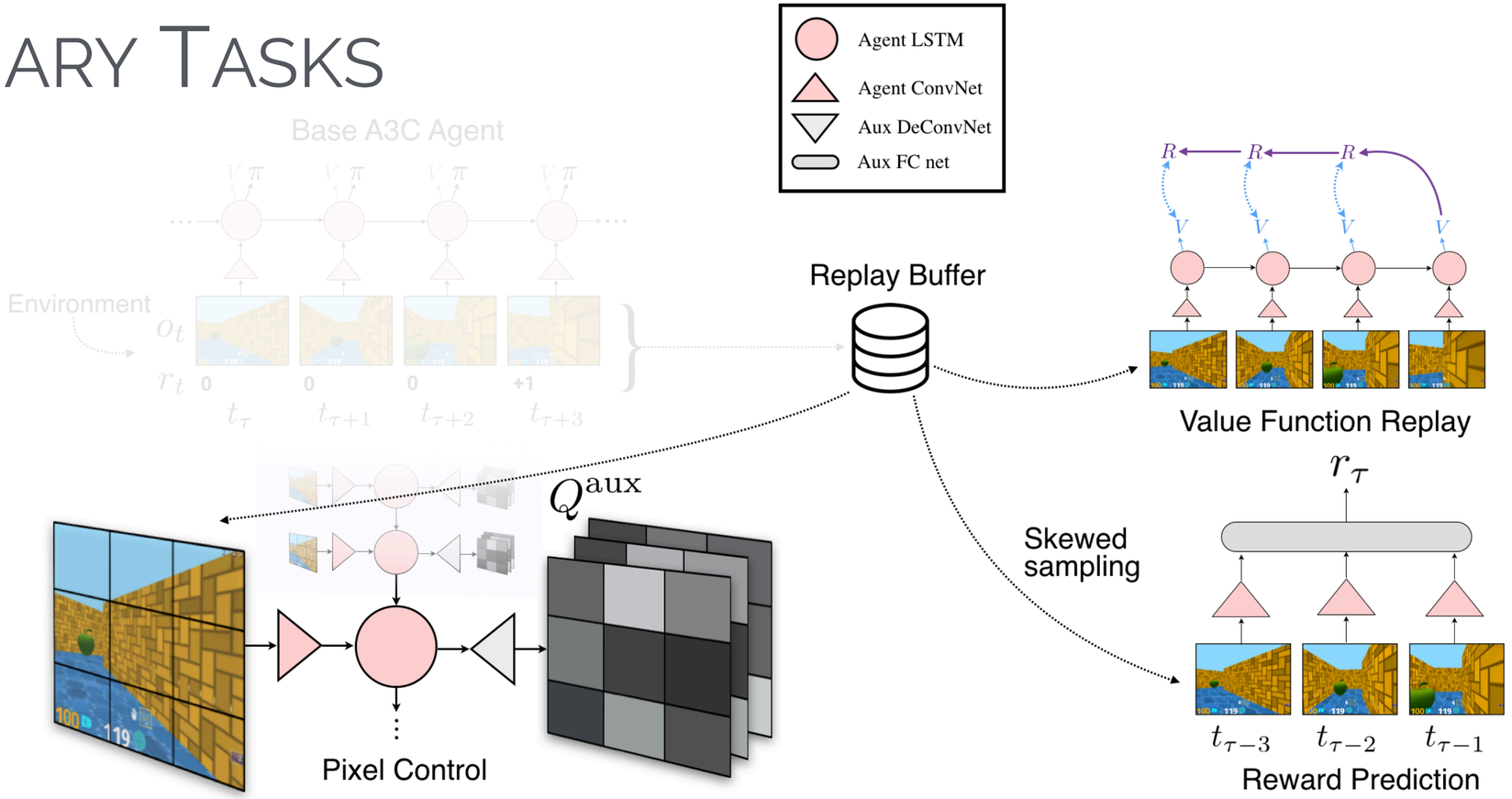
Advantage actor-critic algorithm with multiple asynchronous workers.



UNREAL always acts in the environment with the base A3C policy.

Stores sequences of transitions in a replay buffer.

Off Policy Auxiliary Tasks



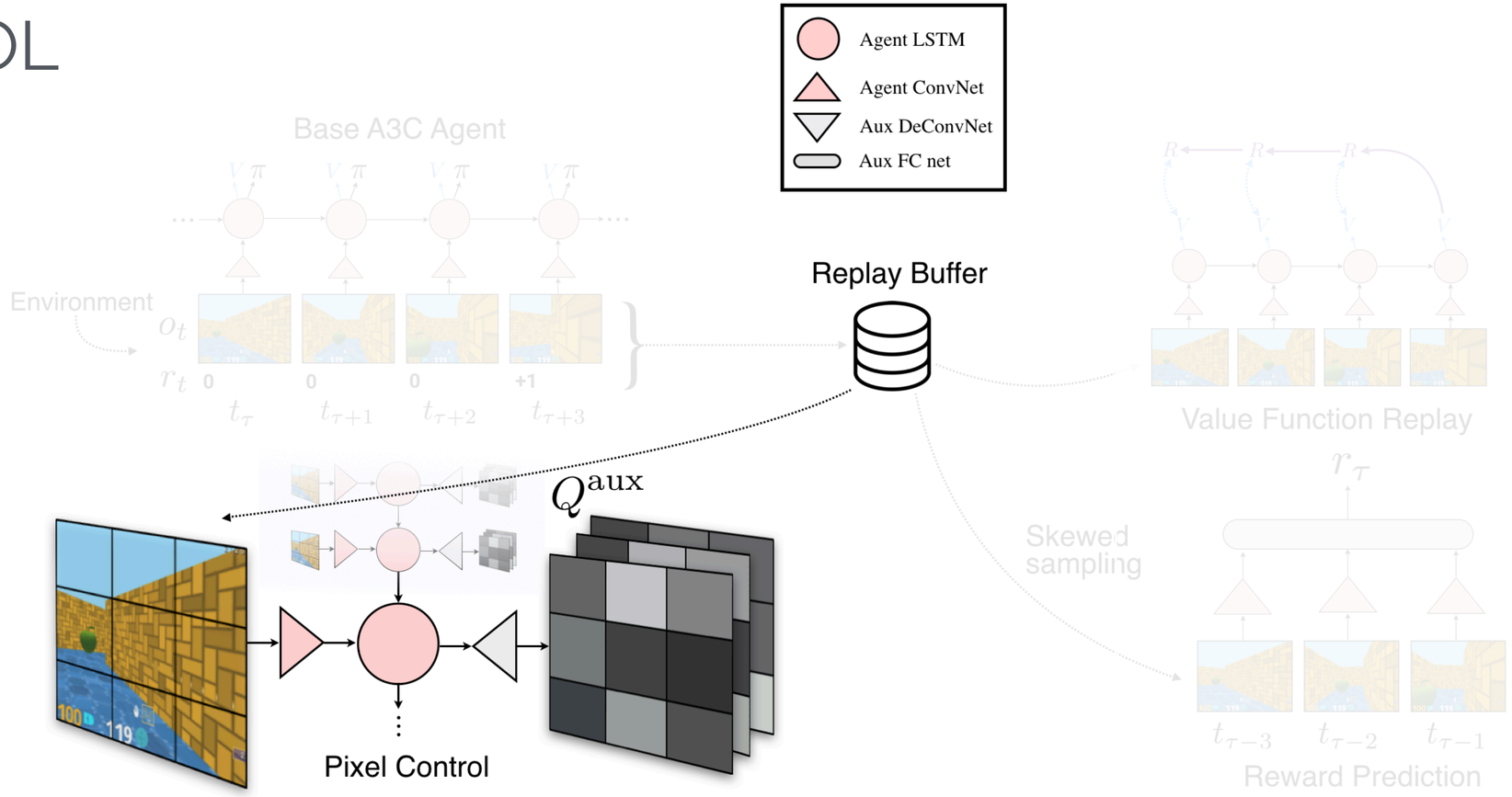
Sampling sequences of transitions from replay allows us to perform auxiliary tasks off-policy **for feature learning** — no need to for off-policy correction.

Networks used for auxiliary tasks are weight shared to agent's network.

Auxiliary Control

Augment A3C with many **auxiliary control tasks**.

Learning to control many aspects of the environment.



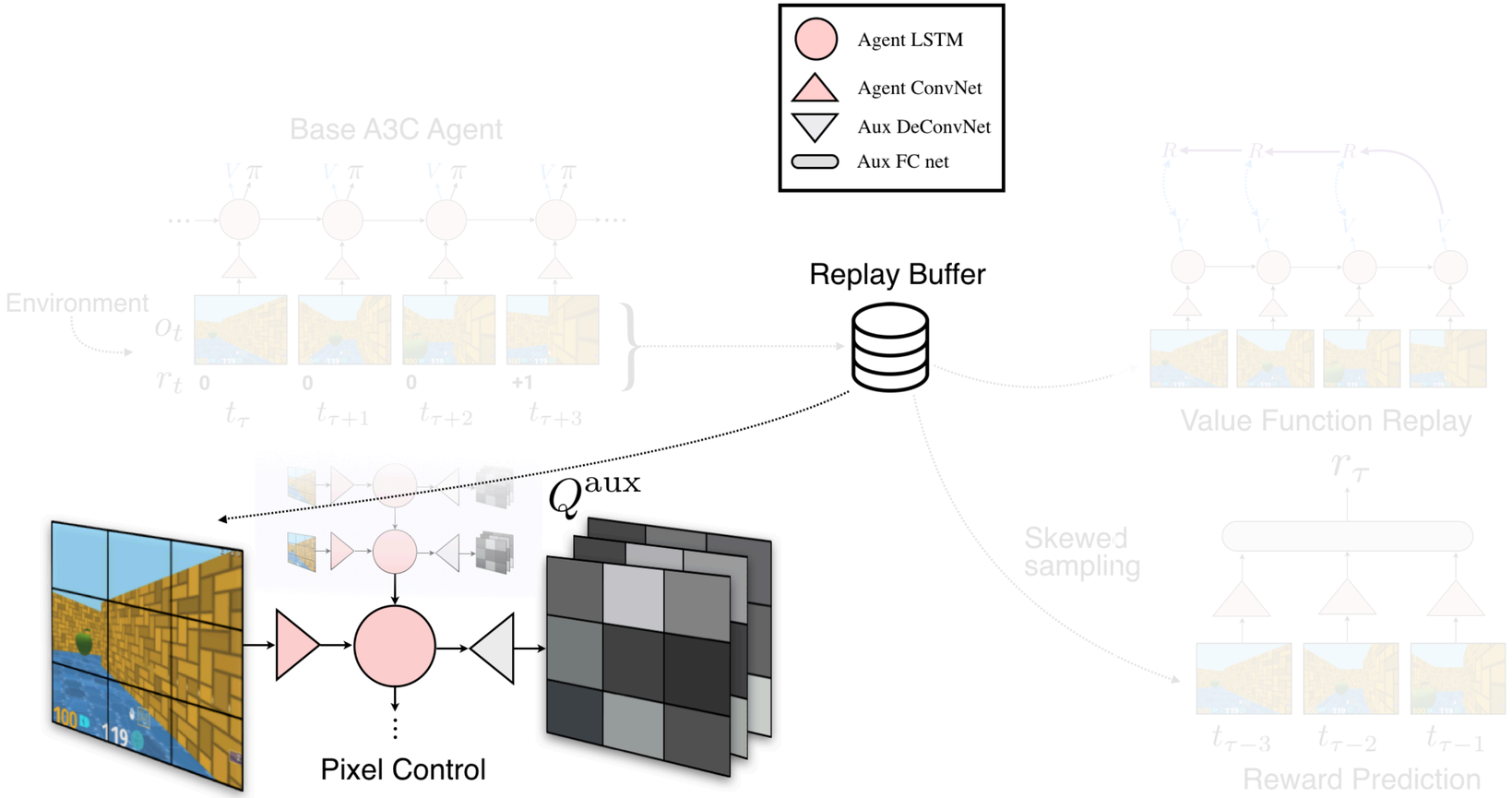
Pixel Control — learn to maximally change parts of the visual input.

Feature Control — learn to control the internal representations.

Pixel Control

Augment A3C with many **auxiliary control tasks**.

Learning to control many aspects of the environment.



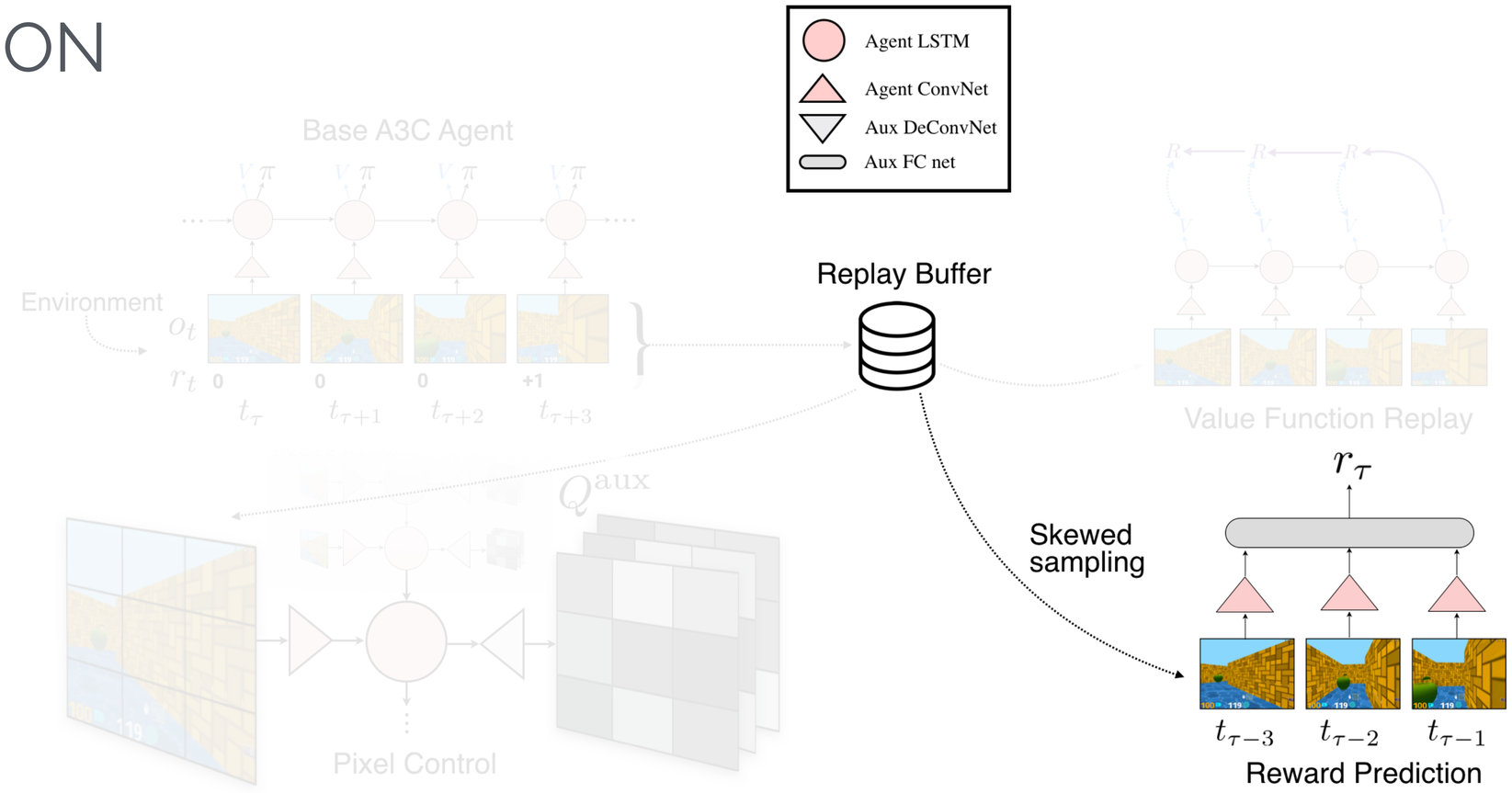
Pixel Control — learn to maximally change parts of the visual input.

- Divide observation into cells.
- Learn a per-cell policy — dueling deconv net *[Wang 2016]*.
- Reward is absolute change in average pixel intensity in cell between time steps.
- Optimise with n-step Q-learning.

Reward Prediction

Focus agent's features on rewarding events.

Shape agent's CNN through prediction task.



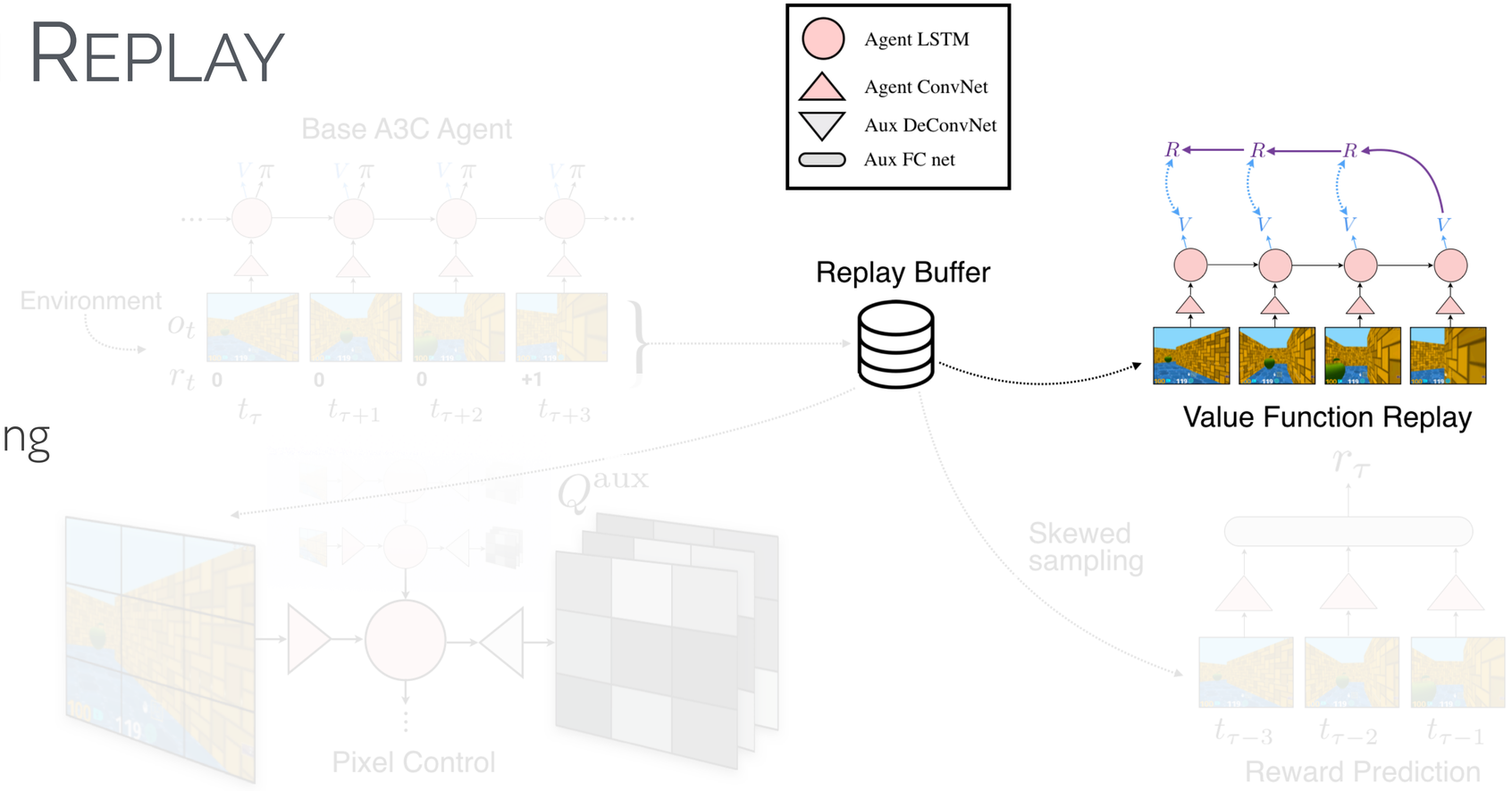
Reward Prediction — learn to classify the next step reward.

- Sample mini-sequences — skewed sampling so 50% sequences end in non-zero reward.
- Encode observations with agent's CNN.
- Concatenated encodings are used to classify +ve, zero, or -ve reward in subsequent frame.

Value Function Replay

Reuse experience to faster train value function.

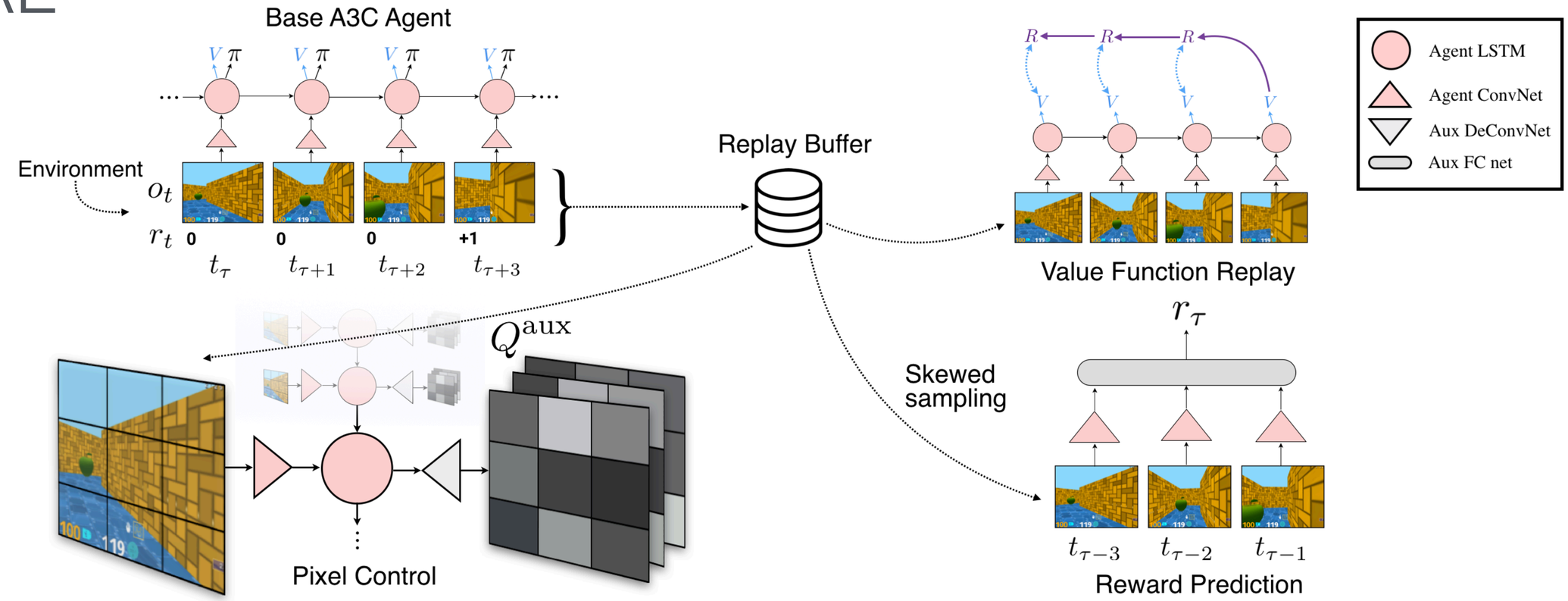
Better features to predict long term reward.



Value Function Replay — further training of value function.

- Sample sequences and perform value regression.
- Bootstraps from values which are not seen on-policy.

UNREAL



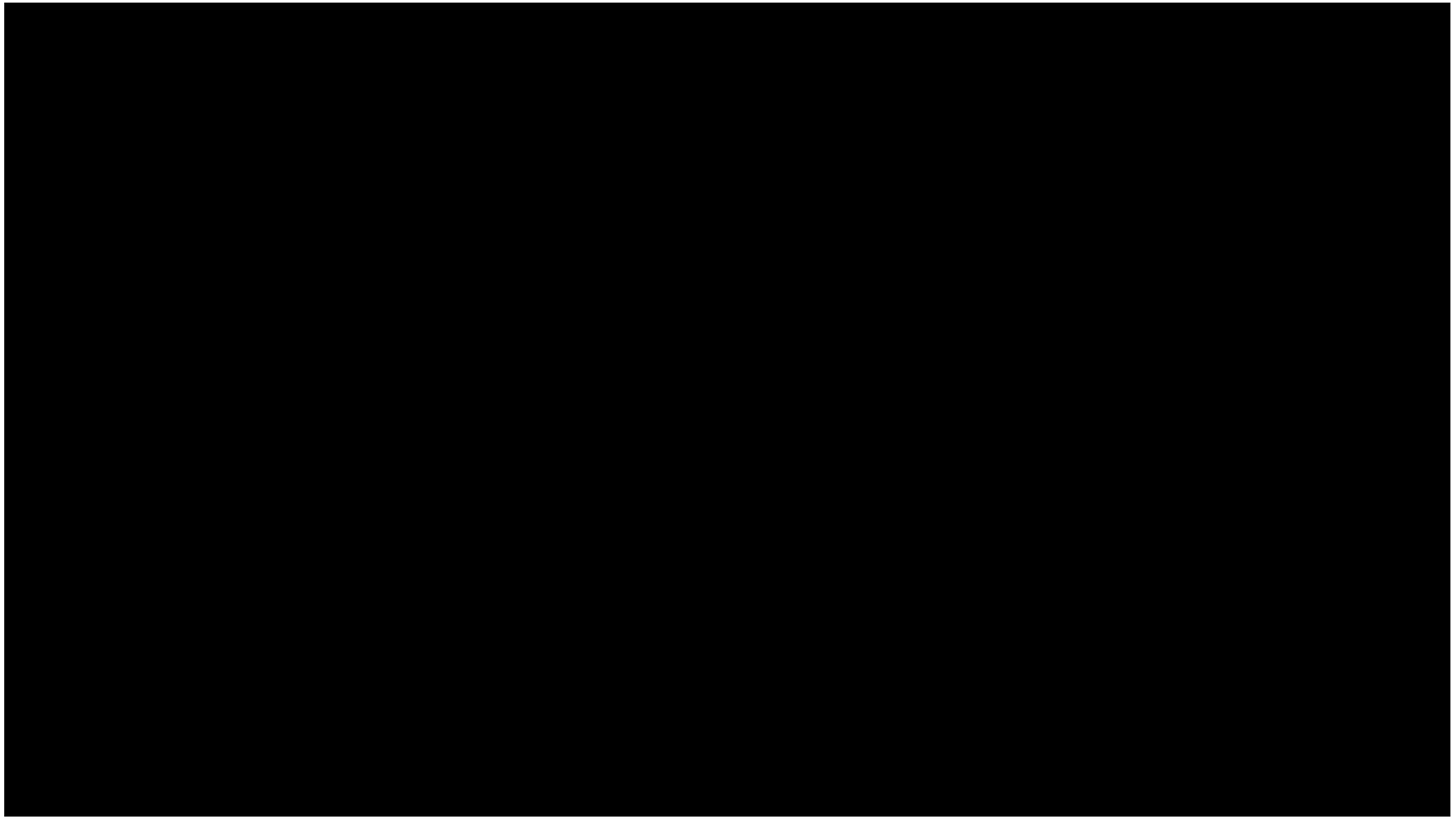
UNREAL optimises all 4 objectives simultaneously.

Minimal tuning on task balancing.

$$\mathcal{L}_{UNREAL}(\theta) = \mathcal{L}_{\text{A3C}} + \lambda_{\text{VR}}\mathcal{L}_{\text{VR}} + \lambda_{\text{PC}} \sum_{c} \mathcal{L}_{Q}^{(c)} + \lambda_{\text{RP}}\mathcal{L}_{\text{RP}}$$

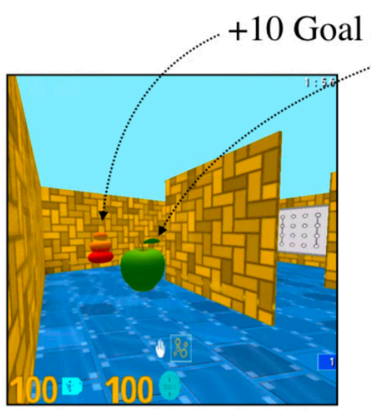

DeepMind Lab

DeepMind Lab EXPERIMENTS

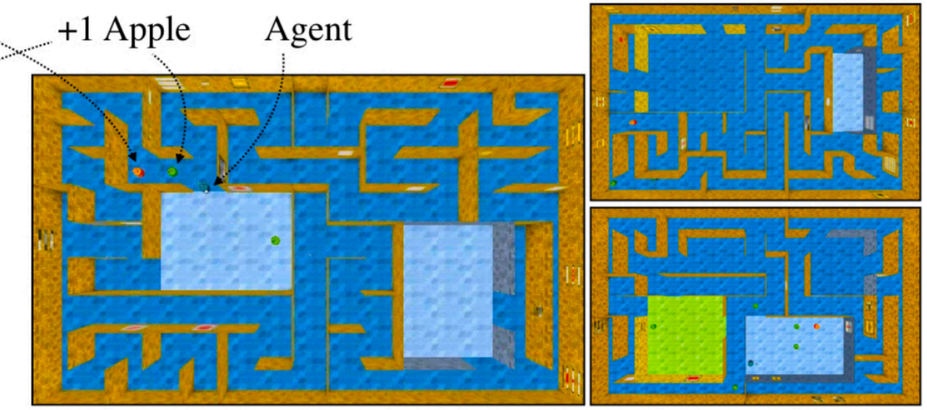


DeepMind Lab EXPERIMENTS

Tested on a suite of 13 levels.
Lasertag, procedural mazes, apple foraging.
Compared to human performance.

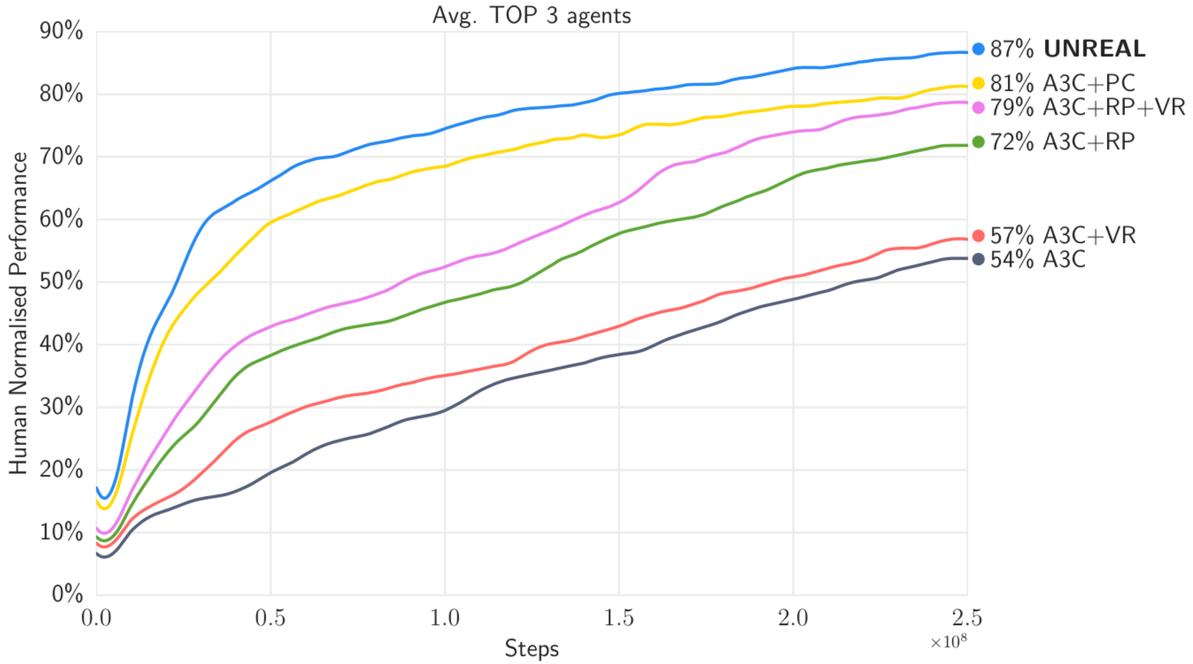


Agent Input



nav_maze_all_random_02 samples

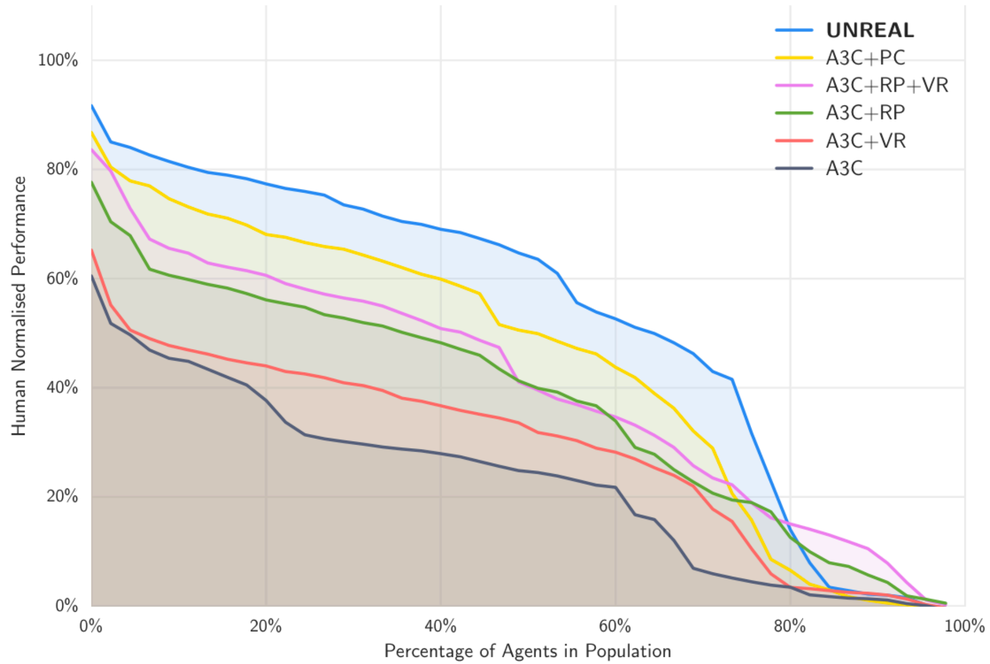
Performance



10x improvement in data efficiency.

60% improvement in final performance.

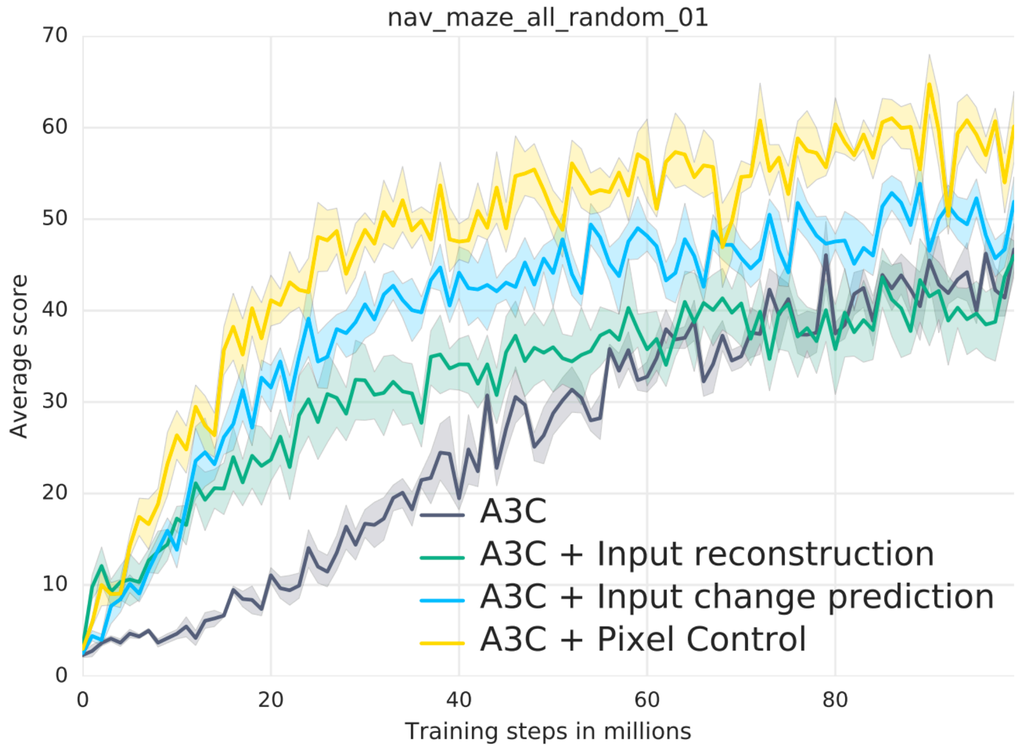
Robustness

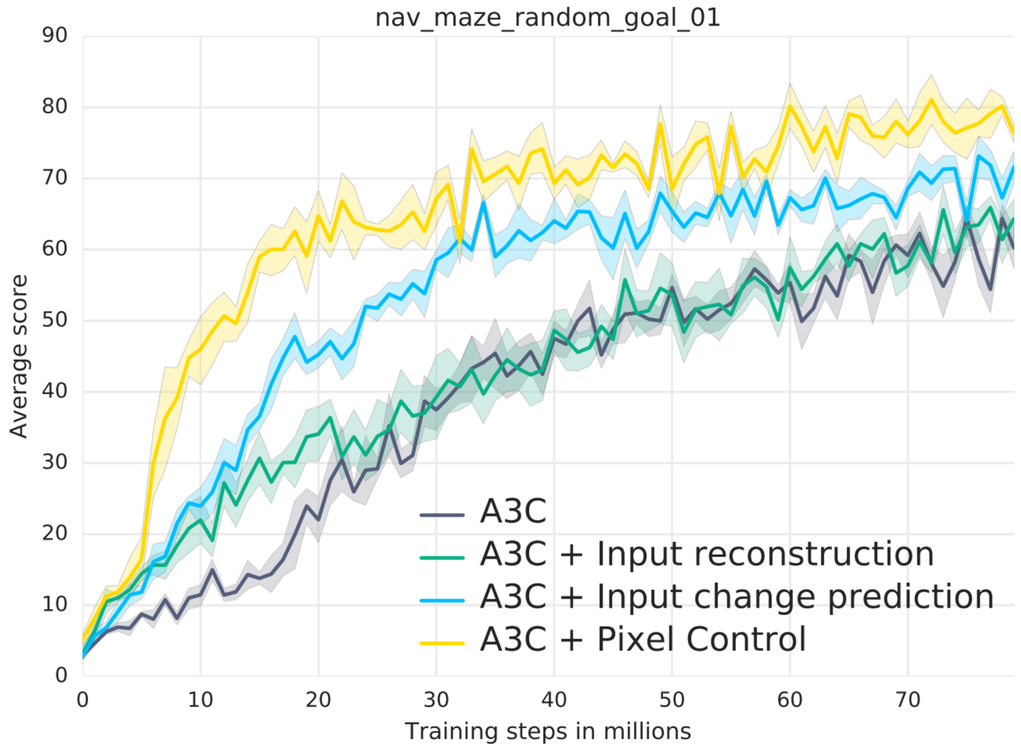


Significantly more robust to hyper parameters.

DeepMind Lab EXPERIMENTS

Control vs prediction vs reconstruction

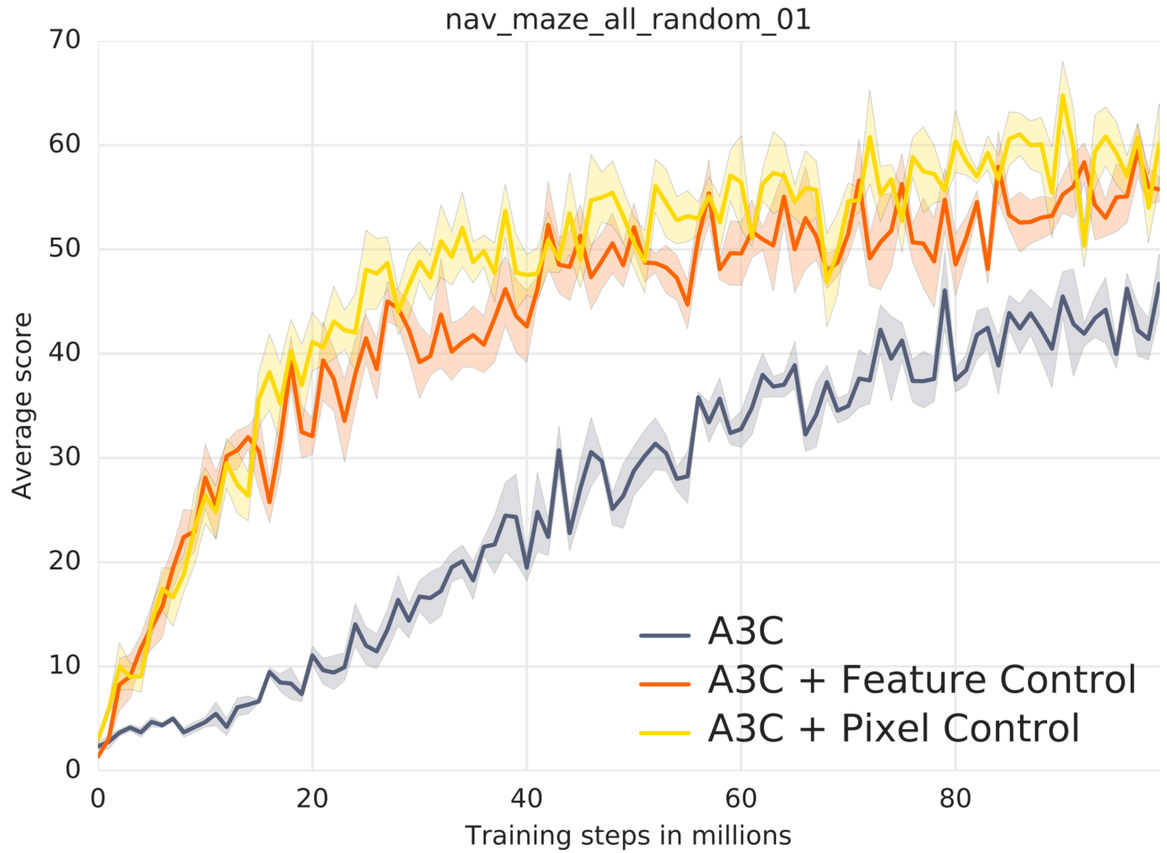




Input **reconstruction** can help initially but often harms policy.
Input **change prediction** helps much more.
Input **control** works best.

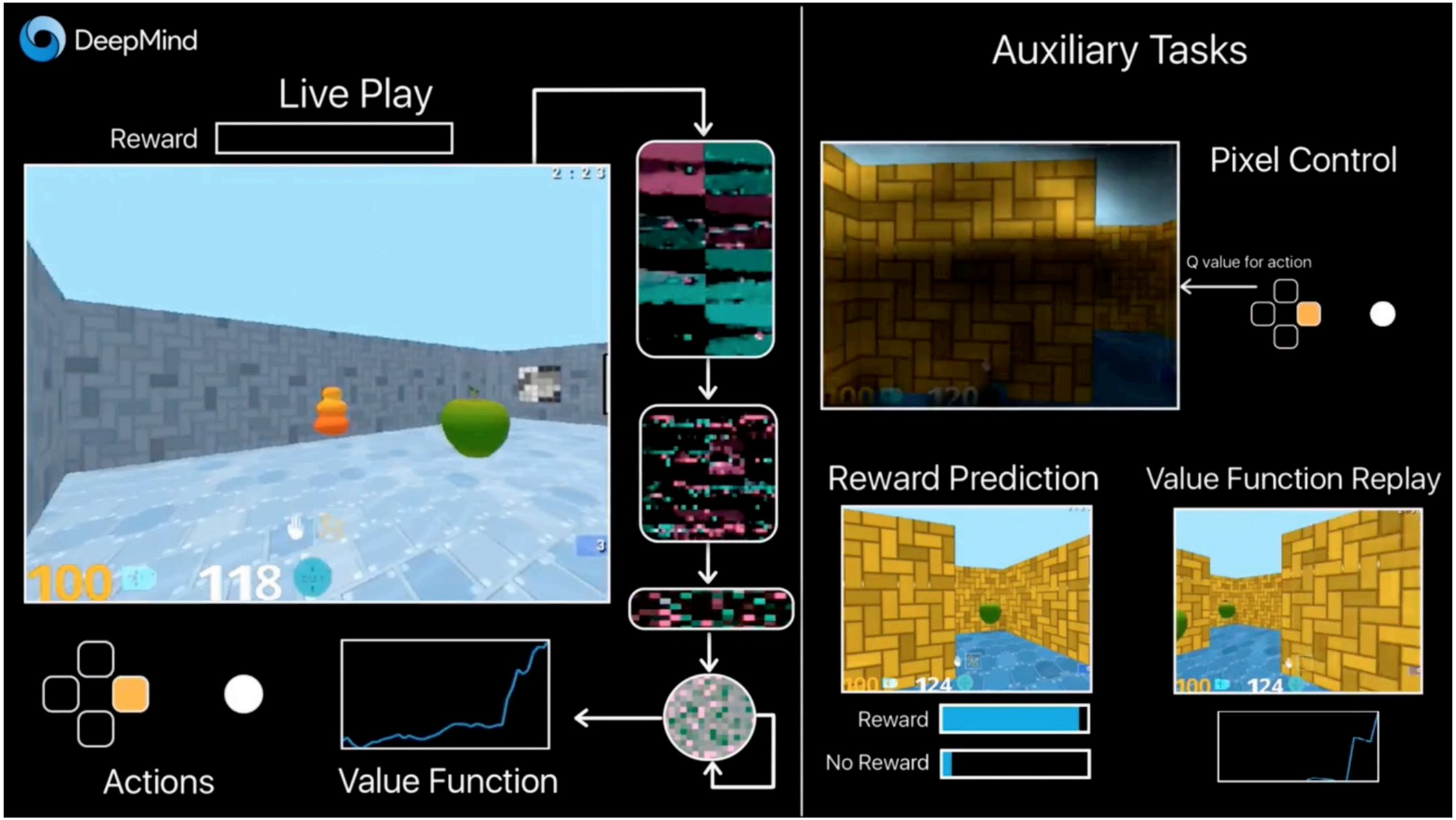
DeepMind Lab EXPERIMENTS

Pixel Control vs Feature Control



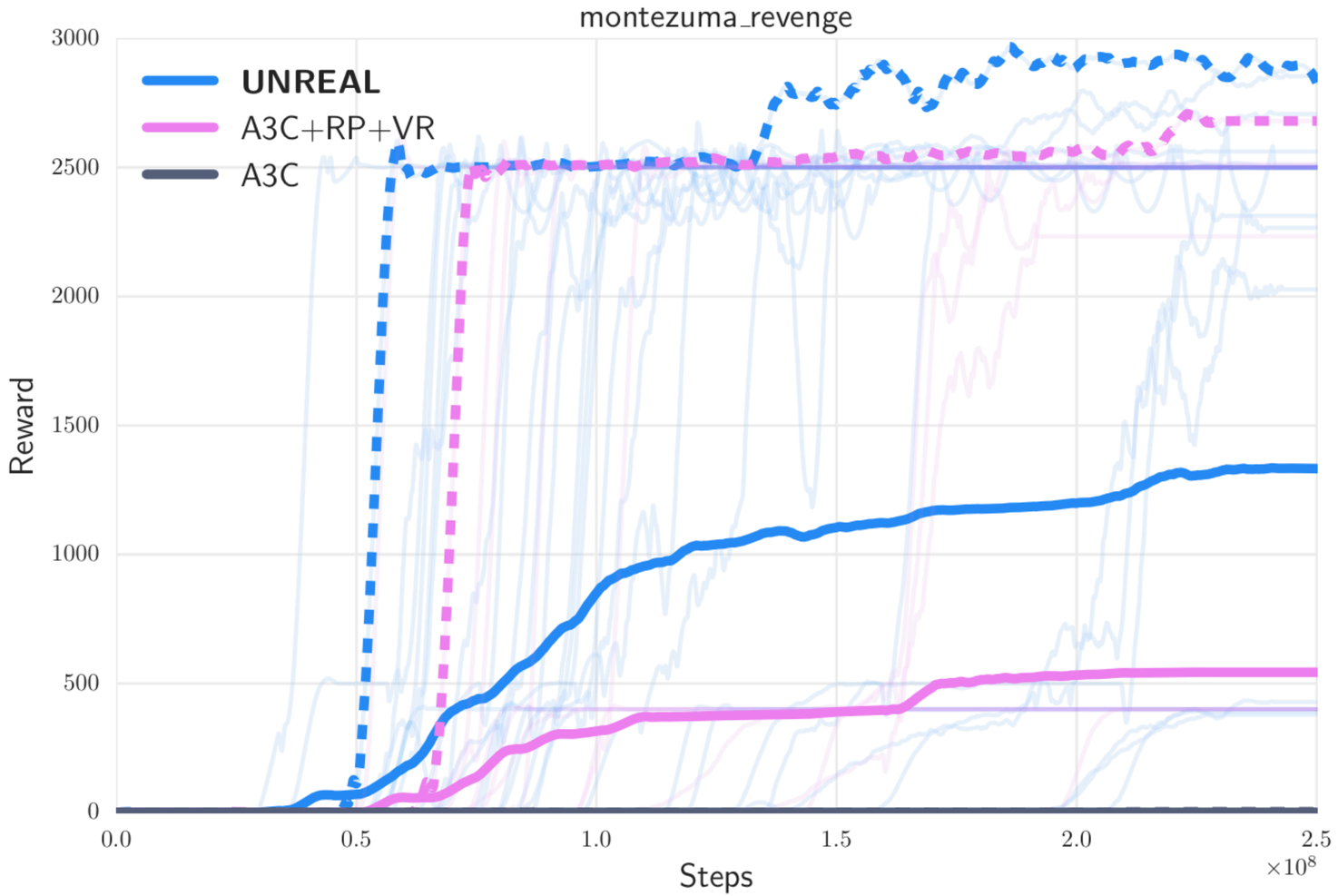
Feature control can work as well as pixel control.
Promising future direction.

DeepMind Lab EXPERIMENTS



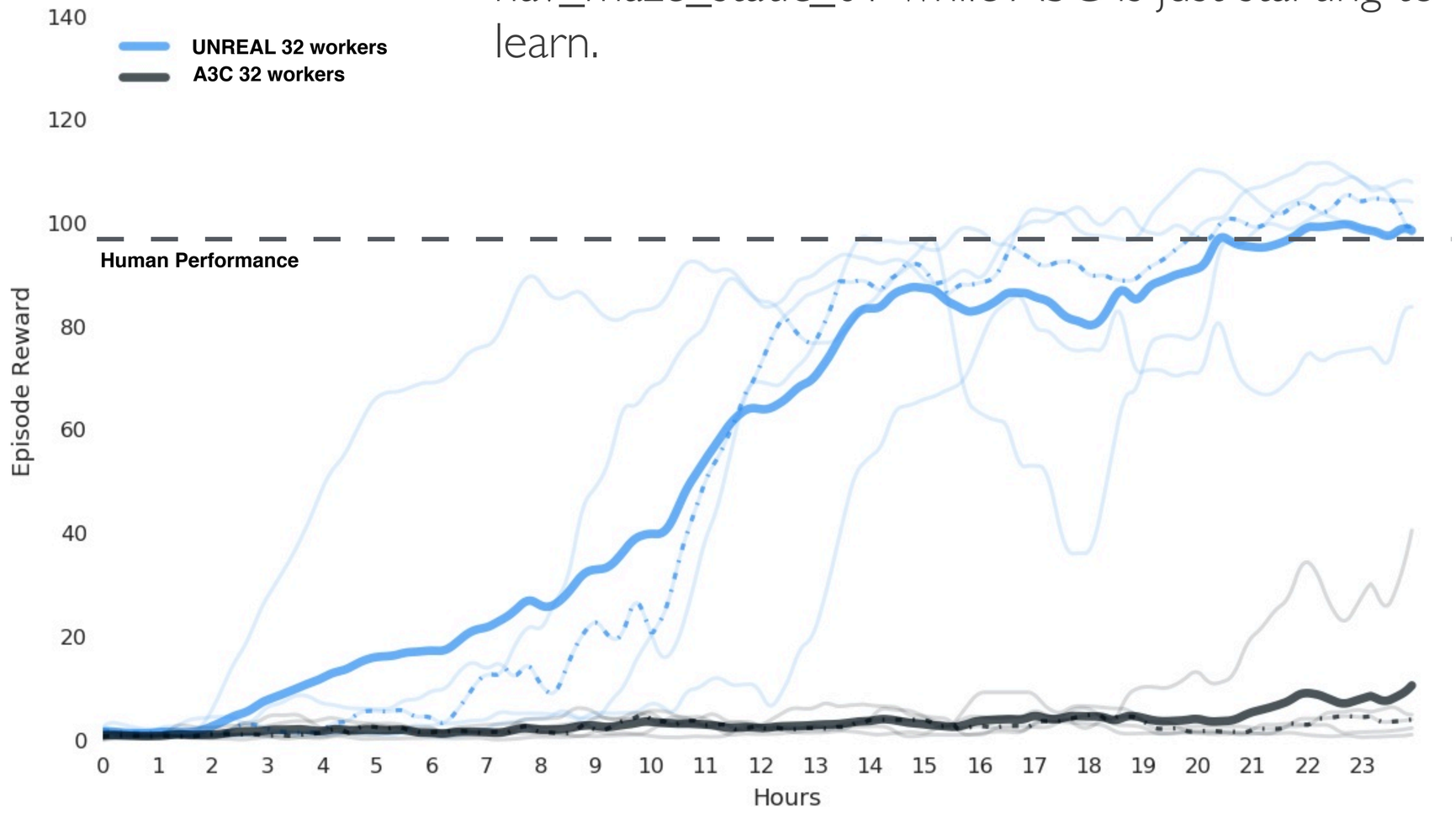
Atari Experiments

State of the art results on Atari:
1453% mean, 331% median of human performance across suite of 57 games.



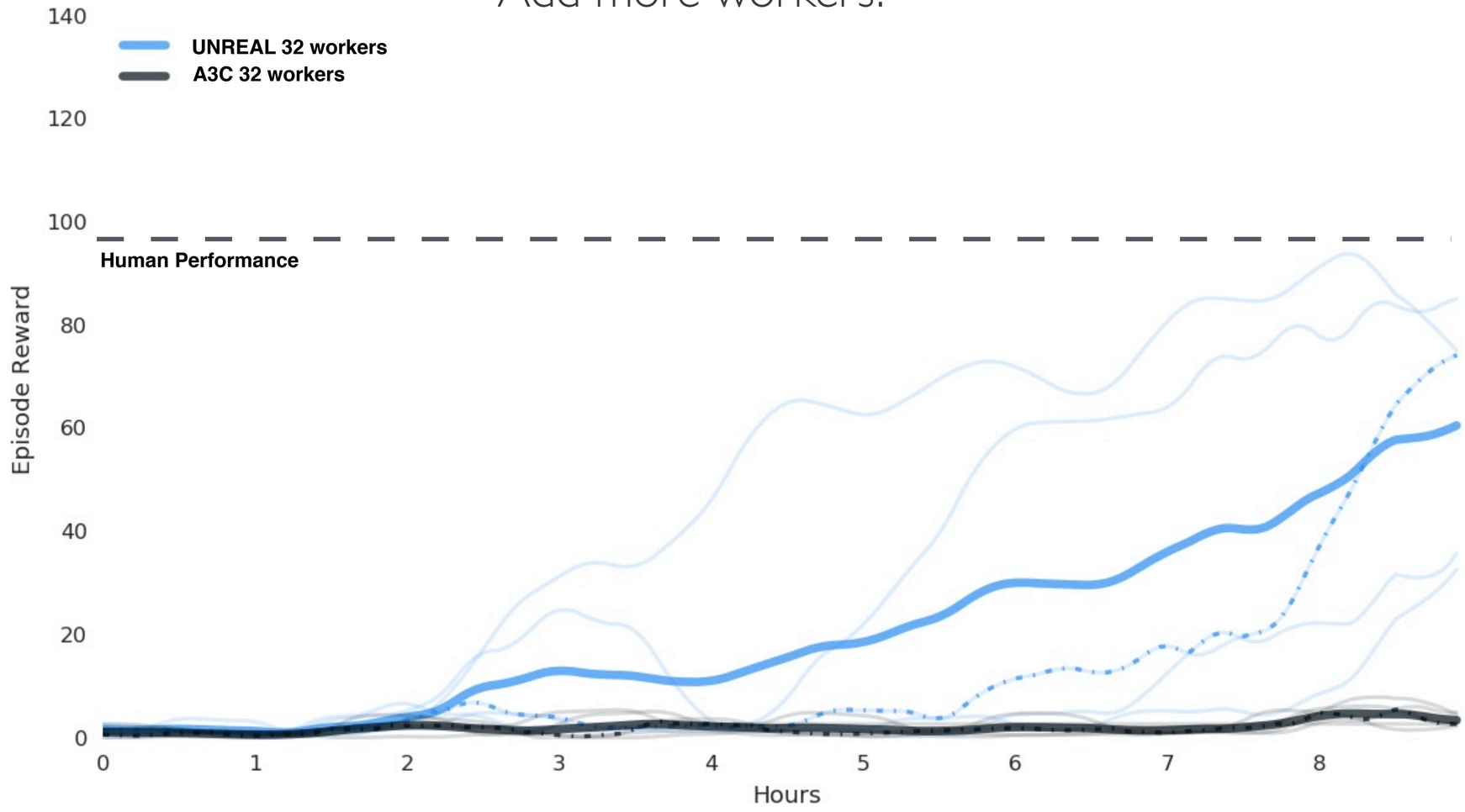
SCALING RL

In 24 hours, UNREAL is close to solving nav_maze_static_01 while A3C is just starting to learn.



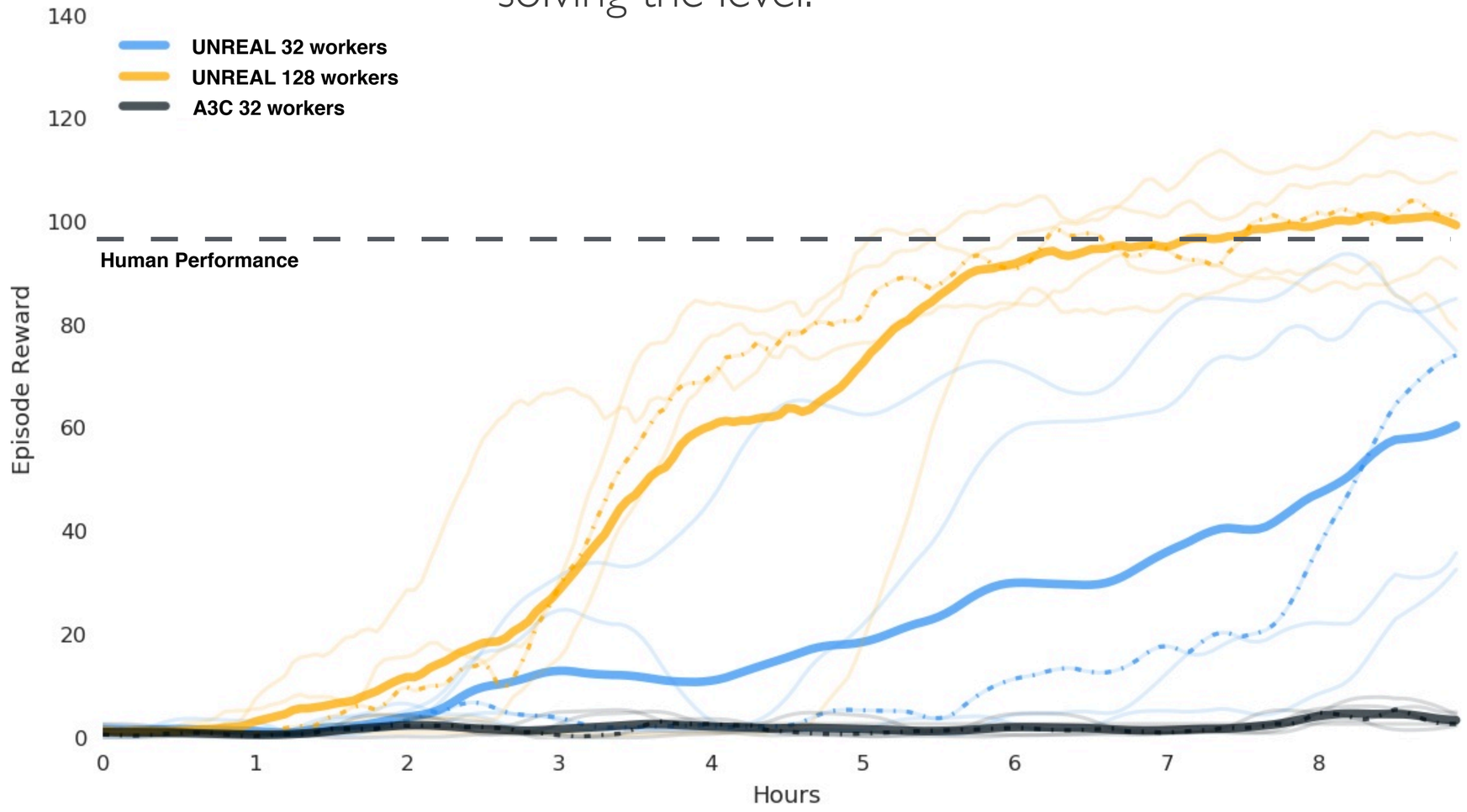
SCALING RL

Overnight, **in 8 hours**, UNREAL starts learning.
Add more workers?



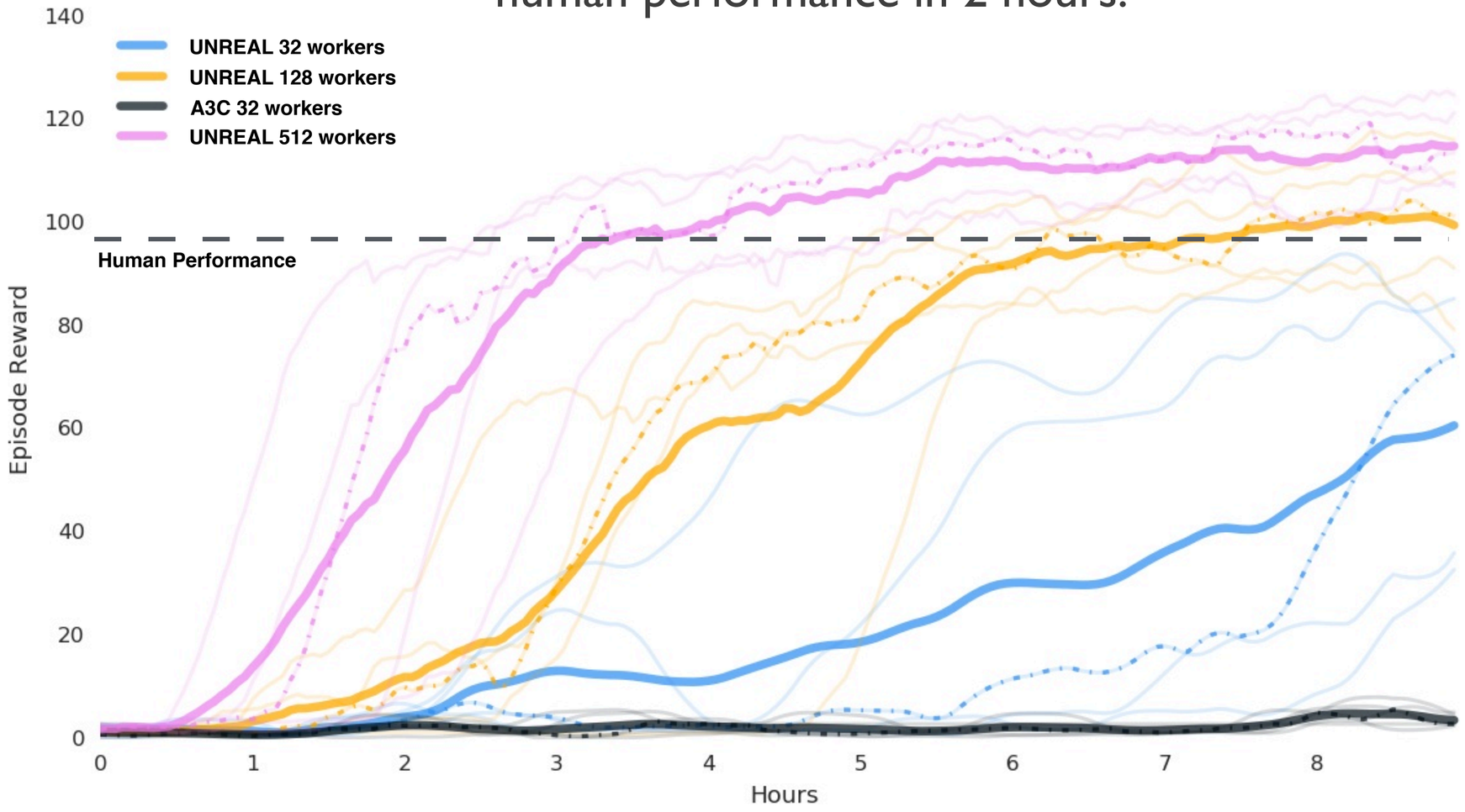
SCALING RL

4x workers means that overnight UNREAL is solving the level.

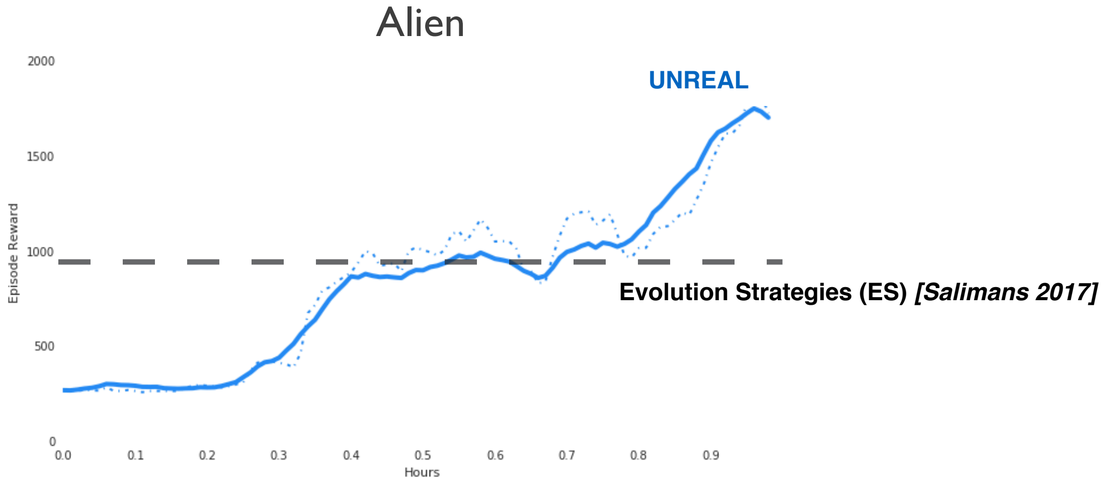


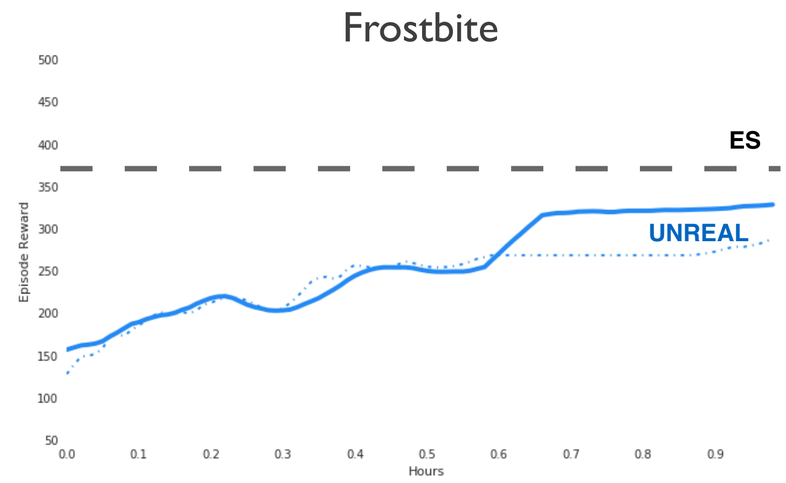
SCALING RL

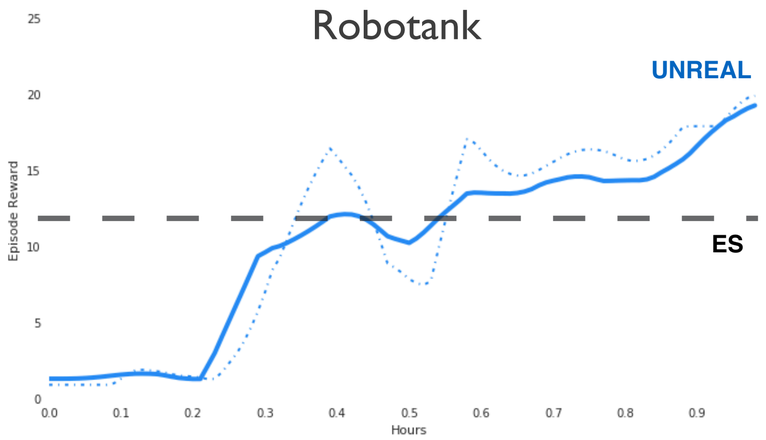
With 512 workers, can reach
human performance in 2 hours.

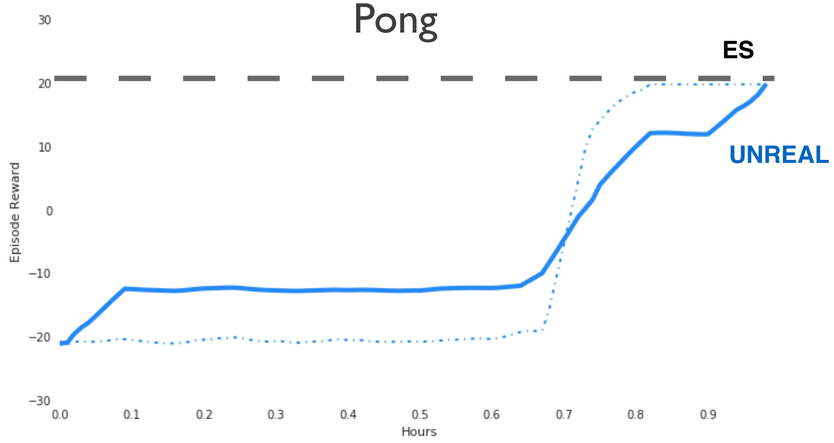


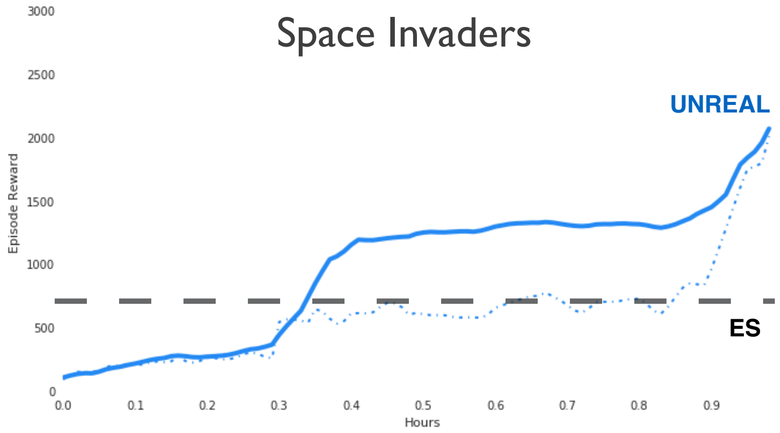
SCALING RL Atari 1k workers











Conclusion

Unsupervised auxiliary tasks drastically improve RL training

- Pixel Control + Reward Prediction + Value Function Replay.
- Improve speed of convergence, robustness to hyperparameters, and results in better performance.
- 10x faster learning and 60% higher performance on DM Lab.
- State of the art results on Atari.

RL scales with CPUs

- Adding more asynchronous workers to A3C-style algorithms greatly speeds up learning.
- Super-human performance on complex domains in a matter of hours.