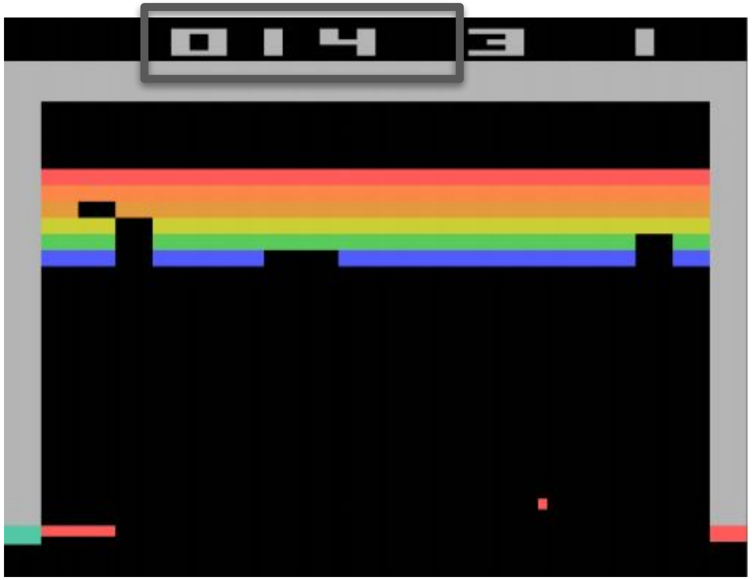
Curiosity-driven Exploration by Self-supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell

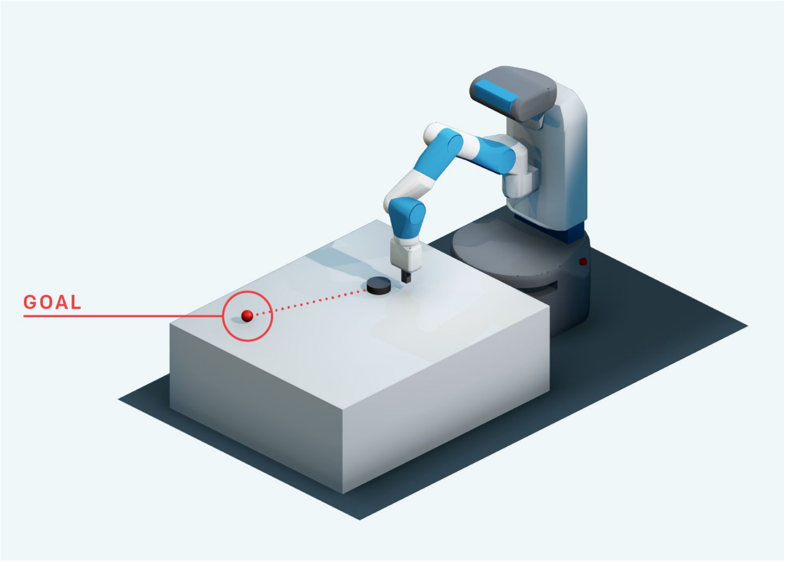
Dense vs Sparse Reward

Dense (ex: Atari)



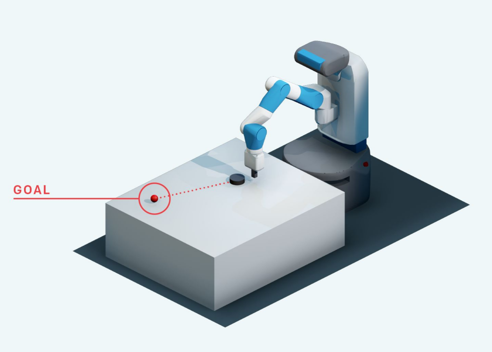
Ex: Reward supplied to the agent at each time step (or frequently, e.g. in some Atari games)

Sparse



Ex: Reward = 1 if agent reaches the goal state, otherwise 0

Dealing with Sparse Reward



- Obtain more **positive/successful trajectory examples** via
  - Using expert **demonstration** (e.g. Imitation learning, Inverse RL)
  - Goal relabeling (e.g. Hindsight Experience Replay (HER))

- **Reward Shaping:** Replace $r_{sparse} \rightarrow r_{shaped}$
  - Engineer shaped reward so that $\pi^*_{sparse} = \pi^*_{shaped}$
    - E.g. Potential-based shaping function (Ng et al, 1999)
  - Augment with **intrinsic reward**: $r_{sparse} + \boxed{r_{intrinsic}}$

Intrinsic Reward

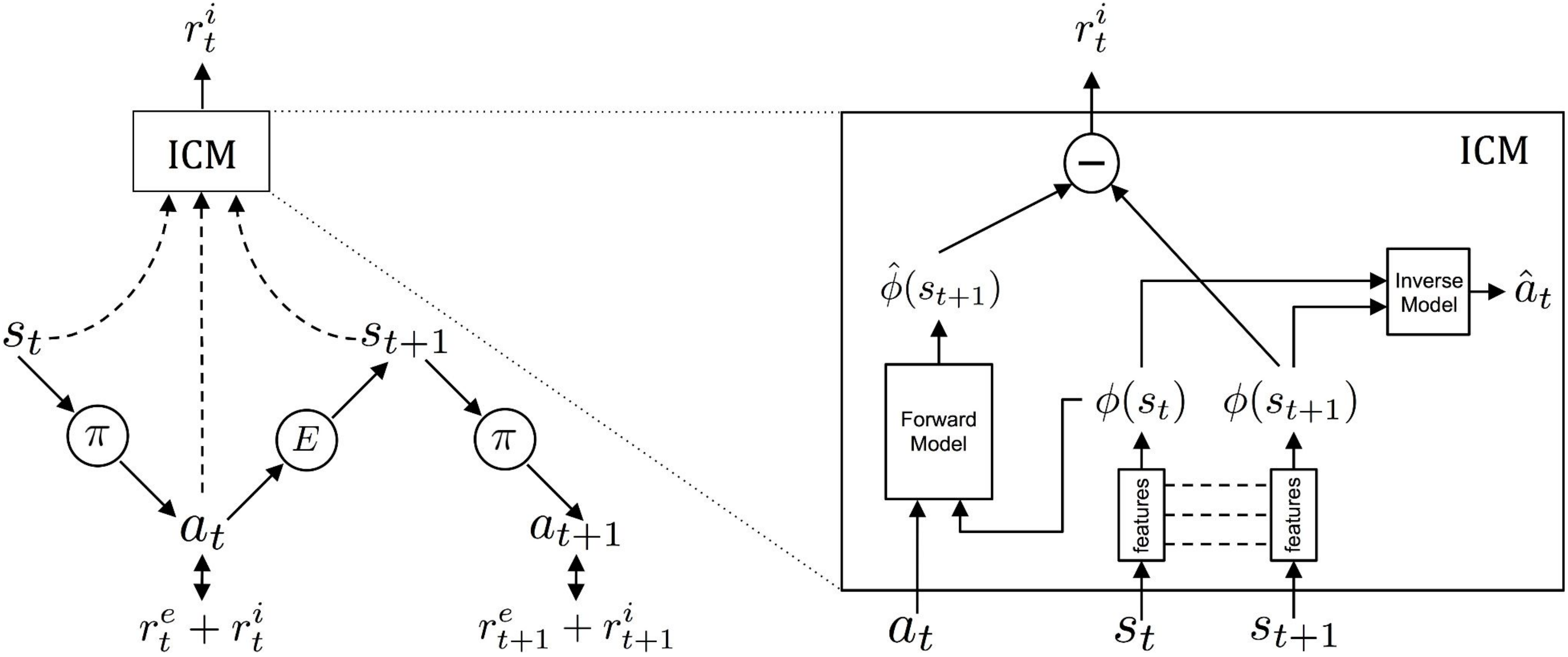
Need $p(s)$

Novelty / Visitation Counts

$p(s_{t+1} | s_t, a_t)$

Prediction Uncertainty / Error

Intrinsic Curiosity Module (ICM)



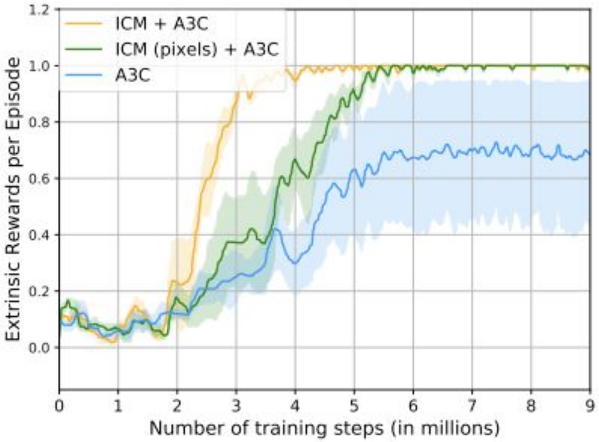
Experiments

Super Mario Bros

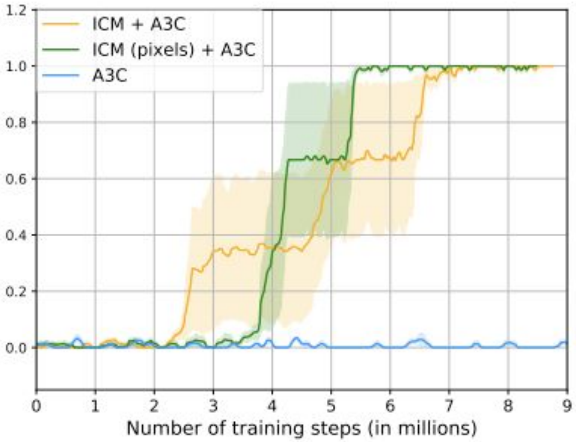
ViZDoom

Use of Curiosity	**Experiment Set Up**
Help Sparse Reward Task	**Sparse** Extrinsic Reward
Help Exploration	**No** Extrinsic Reward
Generalize to novel scenario	**No** extrinsic reward + **new maps**

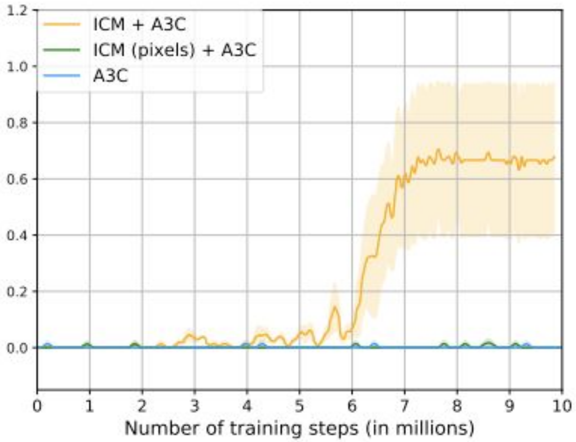
Experiment: Sparse Extrinsic Reward



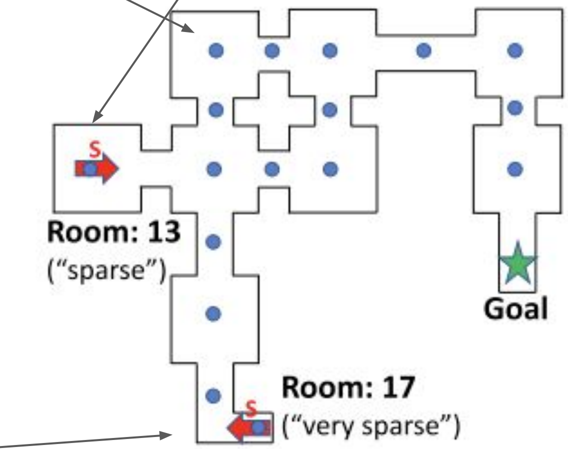
(a) “dense reward” setting



(b) “sparse reward” setting

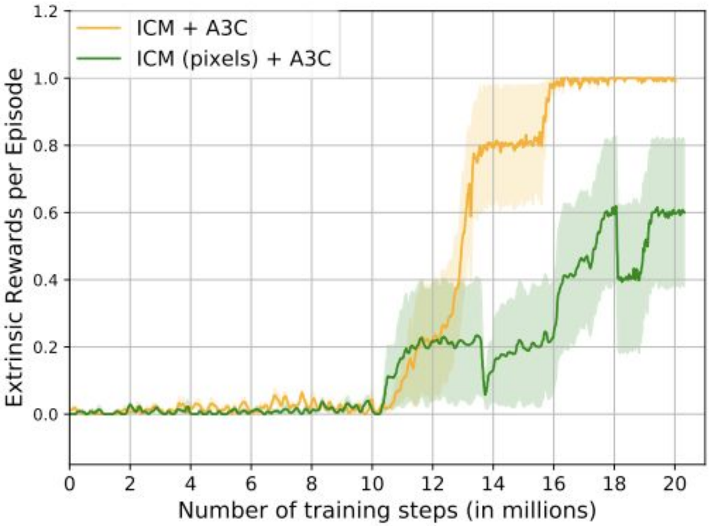


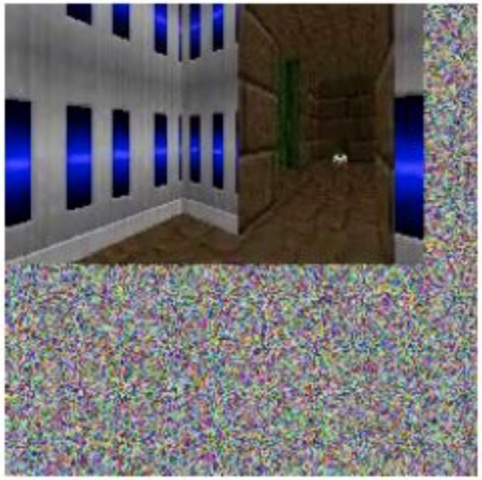
(c) “very sparse reward” setting



- Curious agent is superior to baseline agent w/o curiosity reward (learn faster/at all)
- ICM-Pixels may have trouble with different textures in each room
- ICM better for hard goal directed exploration tasks

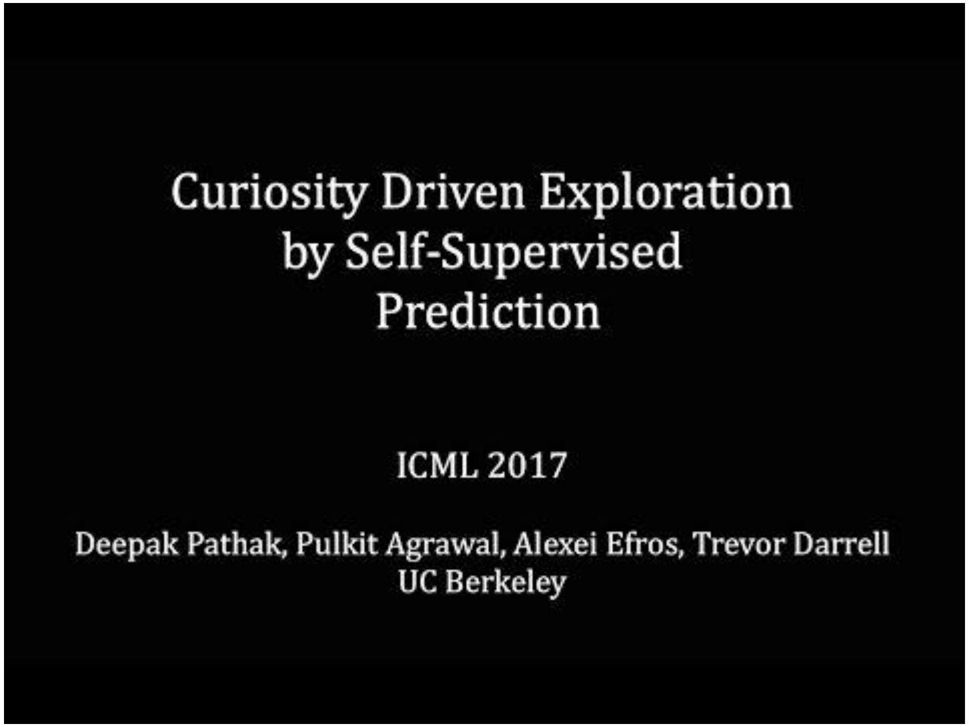
Experiment: Robustness to Uncontrollable Dynamics





(b) Input w/ noise

- Fixed region of white noise (40% of image) as distractors
- “Sparse” reward set up from earlier
- **Result:** ICM solves the task while ICM-pixels suffers



Curiosity Driven Exploration
by Self-Supervised
Prediction

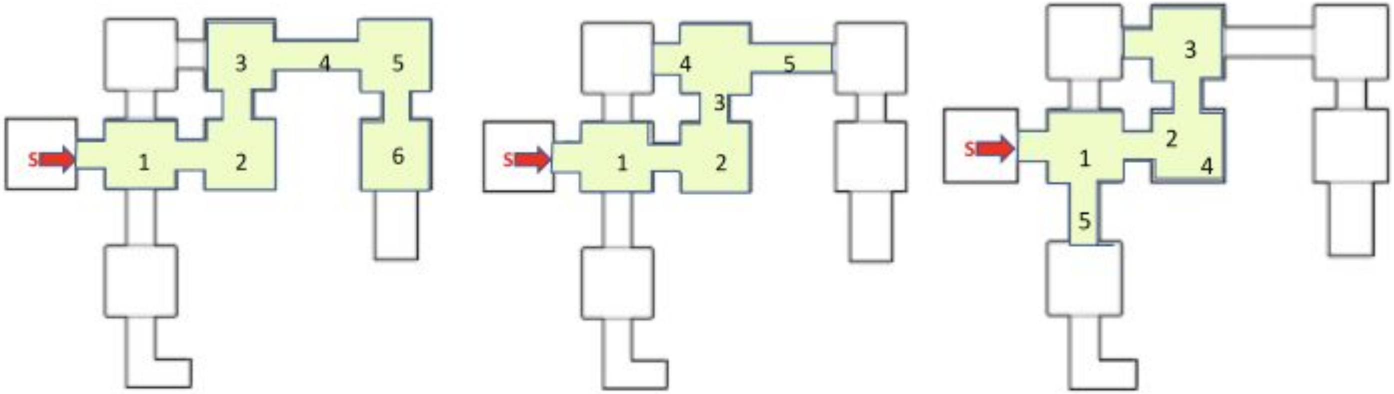
ICML 2017

Deepak Pathak, Pulkit Agrawal, Alexei Efros, Trevor Darrell
UC Berkeley

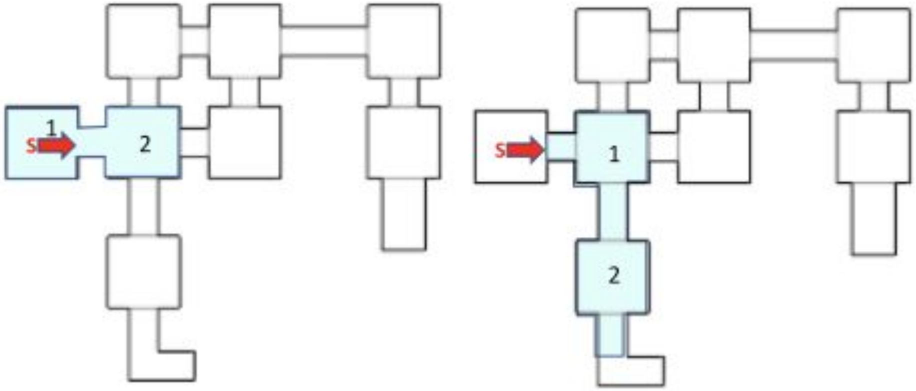
Experiment: No Extrinsic Reward

ViZDoom Visitation pattern during exploration

With Curiosity Reward



Random Action Exploration



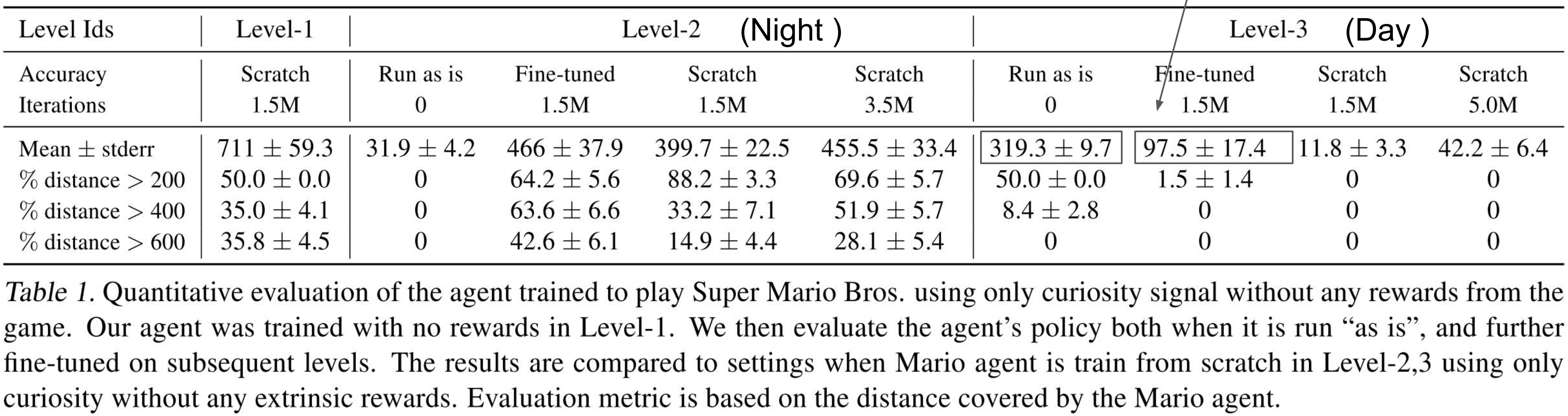
Experiment: Generalization to Novel Scenarios

Train no extrinsic reward, then evaluate:

1. "**As is**" in new scenario
2. **Fine tuning** with *only* curiosity reward
3. Adapt policy to maximize some extrinsic reward

Finetuning causes **degeneracy on Level-3** due to difficult bottleneck point **->** No curiosity reward before the difficult point (familiar)

(Day)



Level Ids	Level-1	Level-2 (Night)				Level-3 (Day)			
Accuracy	Scratch	Run as is	Fine-tuned	Scratch	Scratch	Run as is	Fine-tuned	Scratch	Scratch
Iterations	1.5M	0	1.5M	1.5M	3.5M	0	1.5M	1.5M	5.0M
Mean ± stderr	711 ± 59.3	31.9 ± 4.2	466 ± 37.9	399.7 ± 22.5	455.5 ± 33.4	319.3 ± 9.7	97.5 ± 17.4	11.8 ± 3.3	42.2 ± 6.4
% distance > 200	50.0 ± 0.0	0	64.2 ± 5.6	88.2 ± 3.3	69.6 ± 5.7	50.0 ± 0.0	1.5 ± 1.4	0	0
% distance > 400	35.0 ± 4.1	0	63.6 ± 6.6	33.2 ± 7.1	51.9 ± 5.7	8.4 ± 2.8	0	0	0
% distance > 600	35.8 ± 4.5	0	42.6 ± 6.1	14.9 ± 4.4	28.1 ± 5.4	0	0	0	0

Table 1. Quantitative evaluation of the agent trained to play Super Mario Bros. using only curiosity signal without any rewards from the game. Our agent was trained with no rewards in Level-1. We then evaluate the agent's policy both when it is run "as is", and further fine-tuned on subsequent levels. The results are compared to settings when Mario agent is train from scratch in Level-2,3 using only curiosity without any extrinsic rewards. Evaluation metric is based on the distance covered by the Mario agent.

Experiment: Generalization to Novel Scenarios

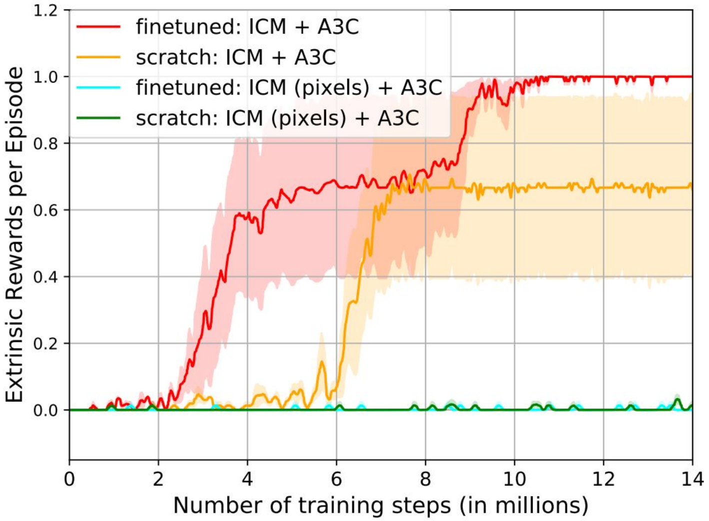


Figure 8. Performance of ICM + A3C agents on the test set of *VizDoom* in the "very sparse" reward case. Fine-tuned models learn the exploration policy without any external rewards on the training maps and are then fine-tuned on the test map. The scratch models are directly trained on the test map. The fine-tuned ICM + A3C significantly outperforms ICM + A3C indicating that our curiosity formulation is able to learn generalizable exploration policies. The pixel prediction based ICM agent completely fail. Note that textures are also different in train and test.

- ICM agent pre-trained only with curiosity + fine-tuned with external reward learns **faster** and achieves **higher reward** than an ICM agent **trained from scratch** to jointly **maximize curiosity and the external rewards**
- **->** Learned exploratory behavior is also useful when the agent is required to achieve goals specified by the environment