

# Curiosity-driven Exploration by Self-supervised Prediction

Петров Тимур, БПМИ-161

# Проблематика

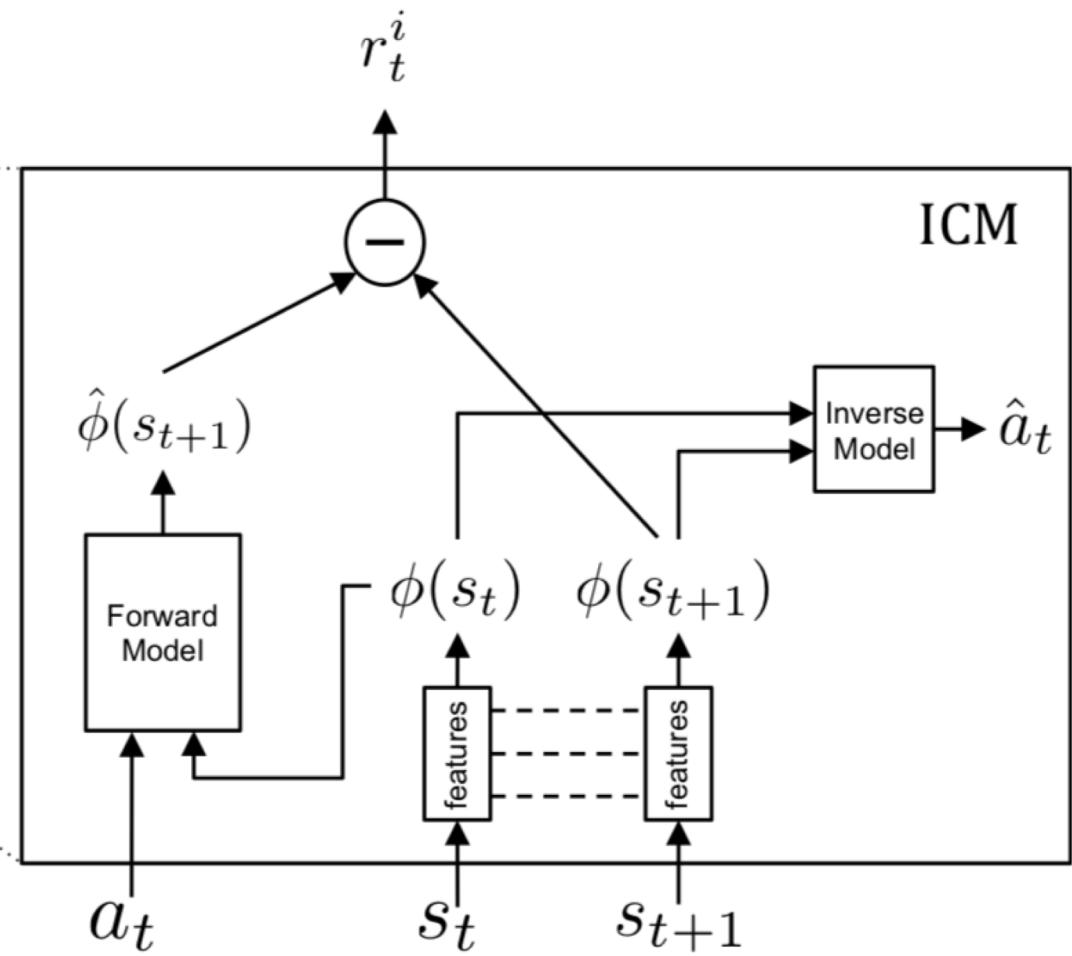
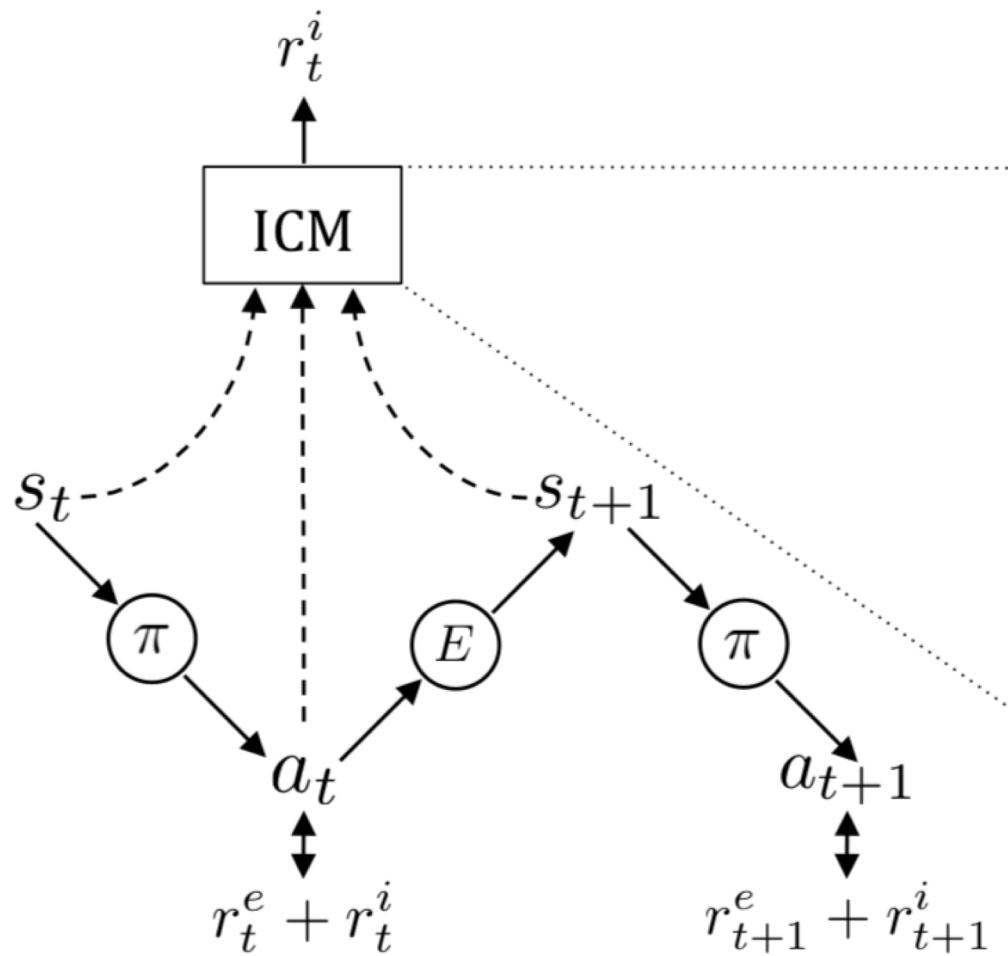
- Большинство решений RL предполагают наличие постоянных наград (прямо или косвенно)
- Но во многих играх и ситуациях наград может или не быть, или же они очень редки (и нет возможности их добавить)
- Идея – добавить любопытство для обучения навыкам
- Для измерения любопытства необходима модель, которую сложно построить

# Любопытство

- Наша награда:  $r_t = r_t^i + r_t^e$ , где  $r_t^i$  – внутренняя награда нашего любопытства, а  $r_t^e$  – внешняя награда (которая в условии редких наград очень часто равна 0)
- Награда за любопытство – ошибка нашего собственного предсказания: чем неожиданней результат, тем больше награда

# Любопытство

- Однако как нам измерять измерять наше любопытство?
- Самая важная деталь: предсказывание изменений только для тех событий, которые зависят от действий нашего агента или же которые влияют на нашего агента, остальное отсекаем



# Модель

- Строим deep-learning модель политики  $\pi(s_t, \theta_p)$ , где параметры  $\theta_p$  подбираются как:

$$\max_{\theta_P} \mathbb{E}_{\pi(s_t, \theta_P)} \left[ \sum_t r_t \right]$$

- Для обучения политики можно выбирать любой метод, вклад является в разработке ICM
- Авторы использовали АЗС (Asynchronous advantage actor critic)

# Inverse Dynamics Model

- Первый подмодуль: обучение пространства признаков  $\phi$ , второй – предсказание действия по данным отображениям:

$$\hat{a}_t = g(s_t, s_{t+1}, \theta_I)$$

$$\min_{\theta_I} L_I(\hat{a}_t, a_t)$$

- Лосс в дискретном случае – это max likelihood для мультиномиального распределения (а выход  $g$  – это softmax)

# Forward Dynamics Model

- Вторая сеть – предсказание результата действия, где

$$\hat{\phi}(s_{t+1}) = f(\phi(s_t), a_t, \theta_F)$$

$$L_F(\phi(s_{t+1}), \hat{\phi}(s_{t+1})) = \frac{1}{2} \left\| \hat{\phi}(s_{t+1}) - \phi(s_{t+1}) \right\|_2^2$$

$$r_t^i = \frac{\eta}{2} \left\| \hat{\phi}(s_{t+1}) - \phi(s_{t+1}) \right\|_2^2$$

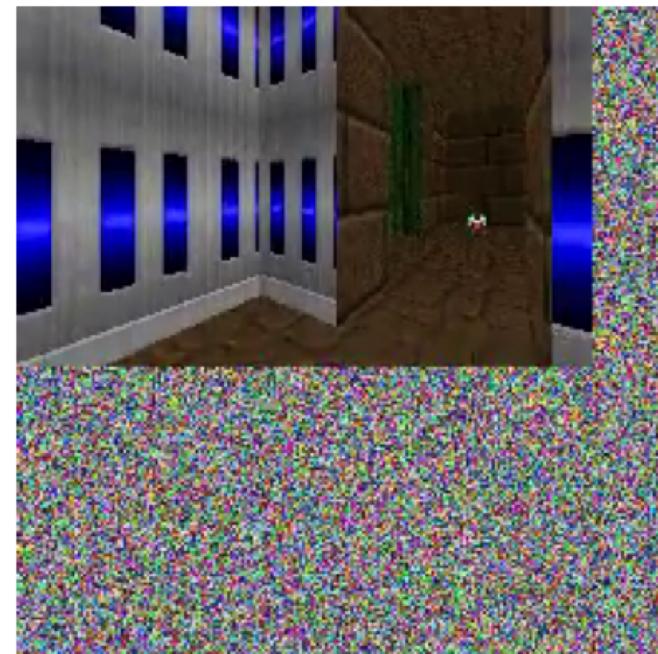
- Итоговое обучение:

$$\min_{\theta_P, \theta_I, \theta_F} \left[ -\lambda \mathbb{E}_{\pi(s_t; \theta_P)} [\Sigma_t r_t] + (1 - \beta) L_I + \beta L_F \right]$$

# Эксперименты (VizDoom)



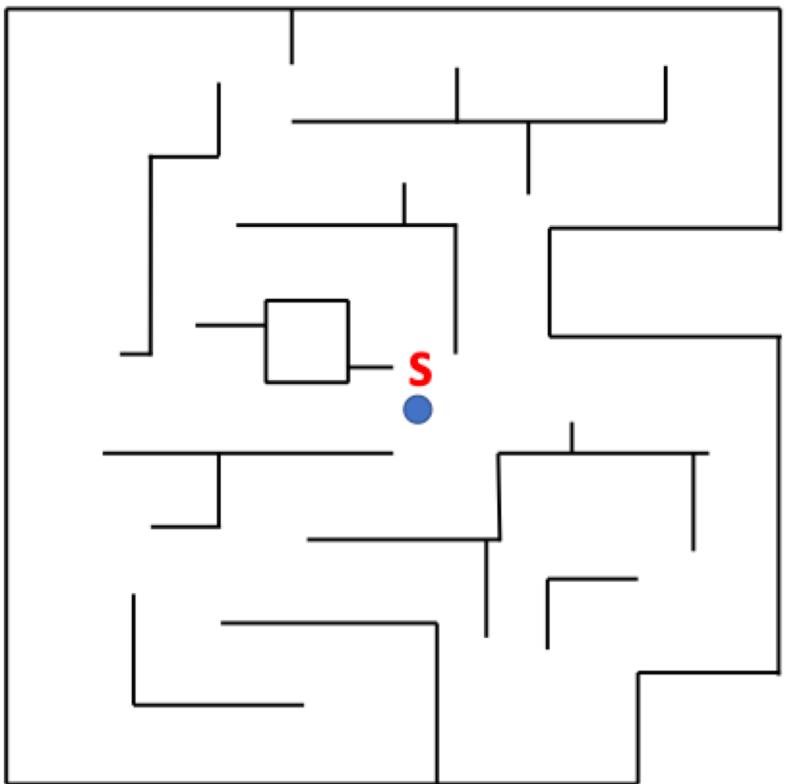
(a) Input snapshot in VizDoom



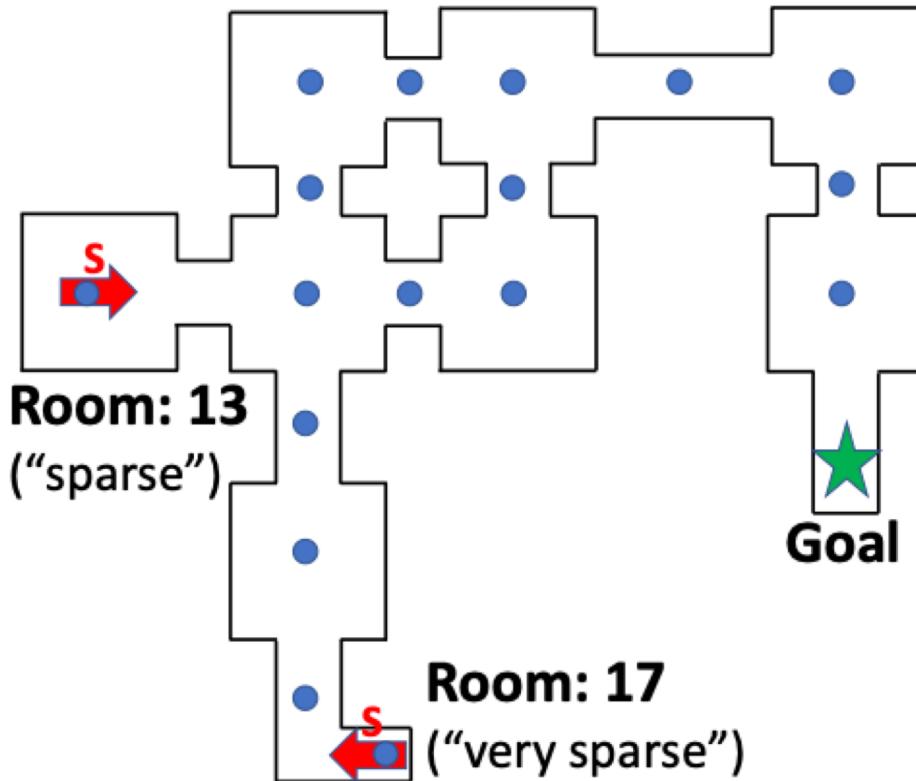
(b) Input w/ noise

*Figure 3.* Frames from VizDoom 3-D environment which agent takes as input: (a) Usual 3-D navigation setup; (b) Setup when uncontrollable noise is added to the input.

# Эксперименты (VizDoom)

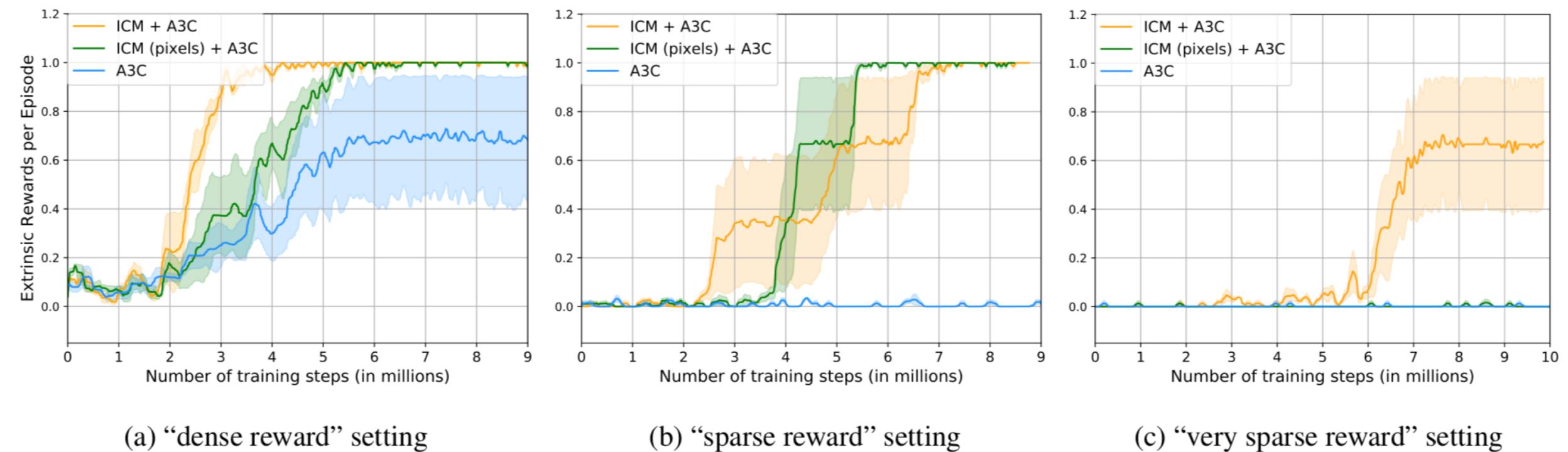


(a) Train Map Scenario

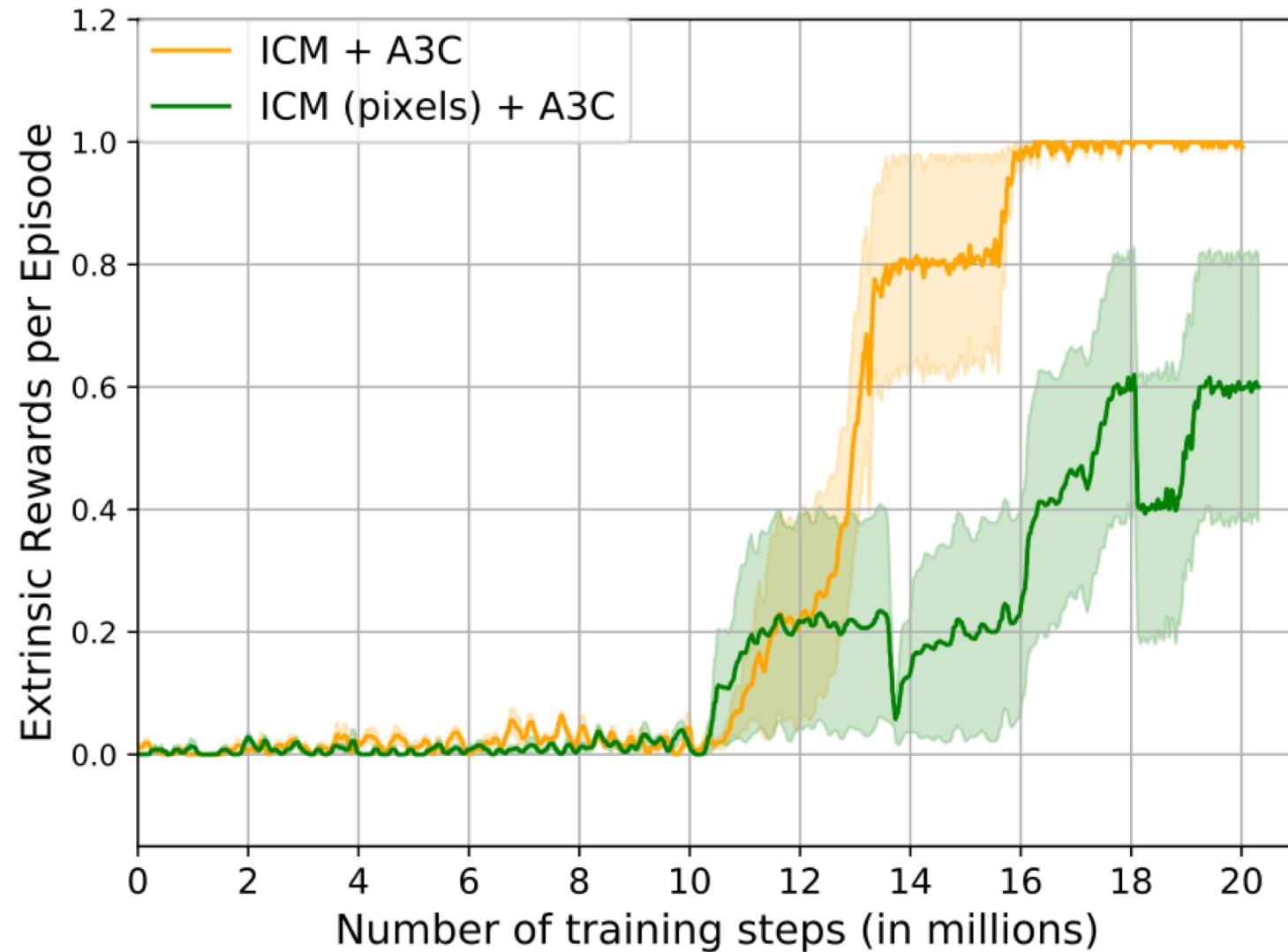


(b) Test Map Scenario

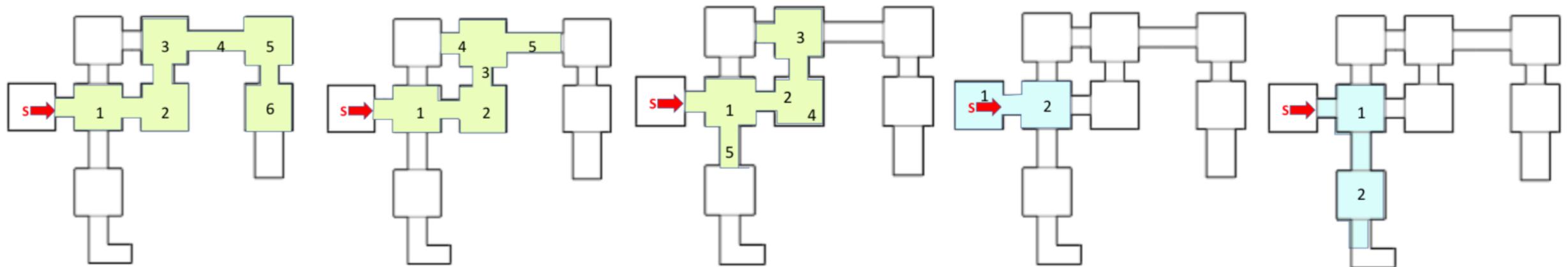
# Результаты (VizDoom)



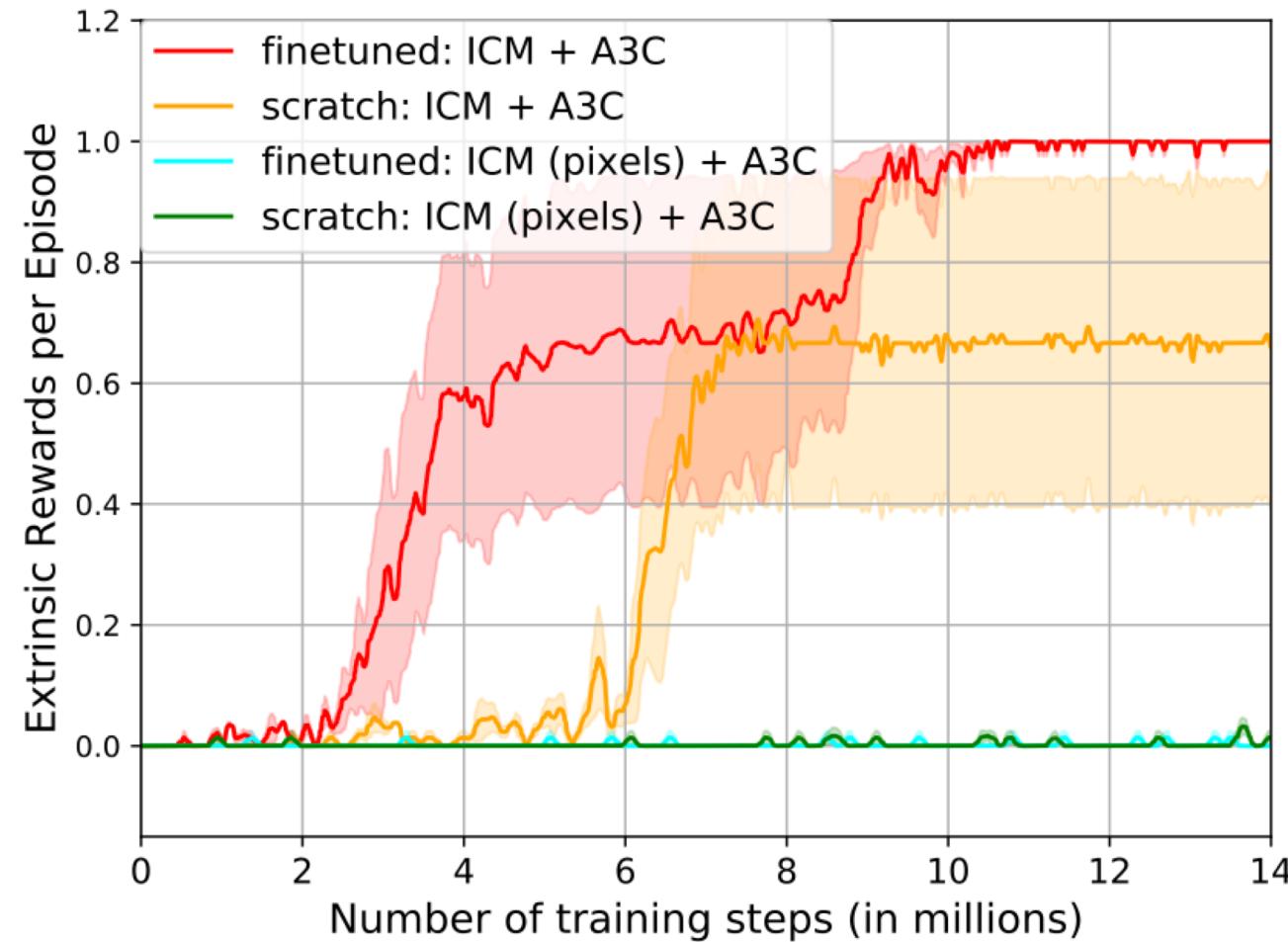
# Влияние “белого шума”



# Ситуация без внешних наград



# Исследование генерализации (VizDoom)



# Результаты (Mario)

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

# Выводы

- Любопытство позволяет исследовать пространство, чтобы выучить методы, которые помогут в будущем
- Предложенная модель ICM позволяет выделить значимые признаки, влияющие на агента, а также ввести награду за любопытство
- Модель позволяет исследовать гораздо больше, а также имеет свойство генерализации накопленного опыта

# Ссылки

- Curiosity-driven Exploration by Self-supervised Prediction,  
<https://arxiv.org/abs/1705.05363.pdf>
- Asynchronous Methods for Deep Reinforcement Learning,  
<https://arxiv.org/pdf/1602.01783.pdf>

# Вопросы

1. Какую проблему решают авторы статьи для задач RL? Какое нововведение они предлагают?
2. Опишите модель ICM. Для чего нужны Forward и Inverse части модели?
3. Выпишите функционал, который используемая модель оптимизирует, поясните каждую из ее частей.