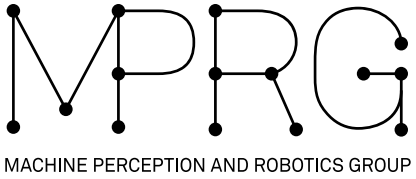
IJCAI2019 Scaling-Up Reinforcement Learning (SURL) Workshop

Adaptive Selection of Auxiliary Tasks in UNREAL

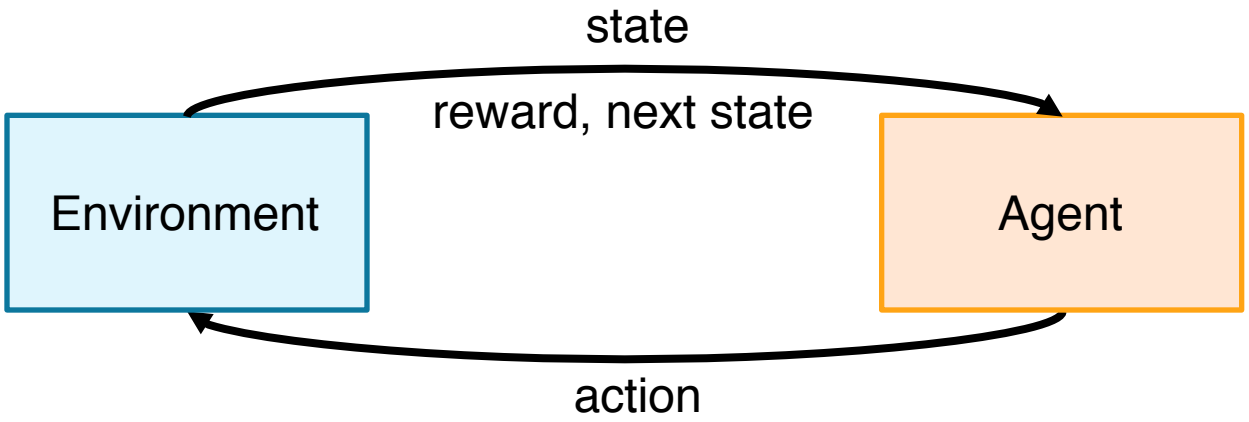
Hidenori Itaya, Tsubasa Hirakawa
Takayoshi Yamashita, Hironobu Fujiyoshi
（Chubu University）



Reinforcement Learning (RL)

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

- Problems involving an agent interacting with an environment



- Application example of RL

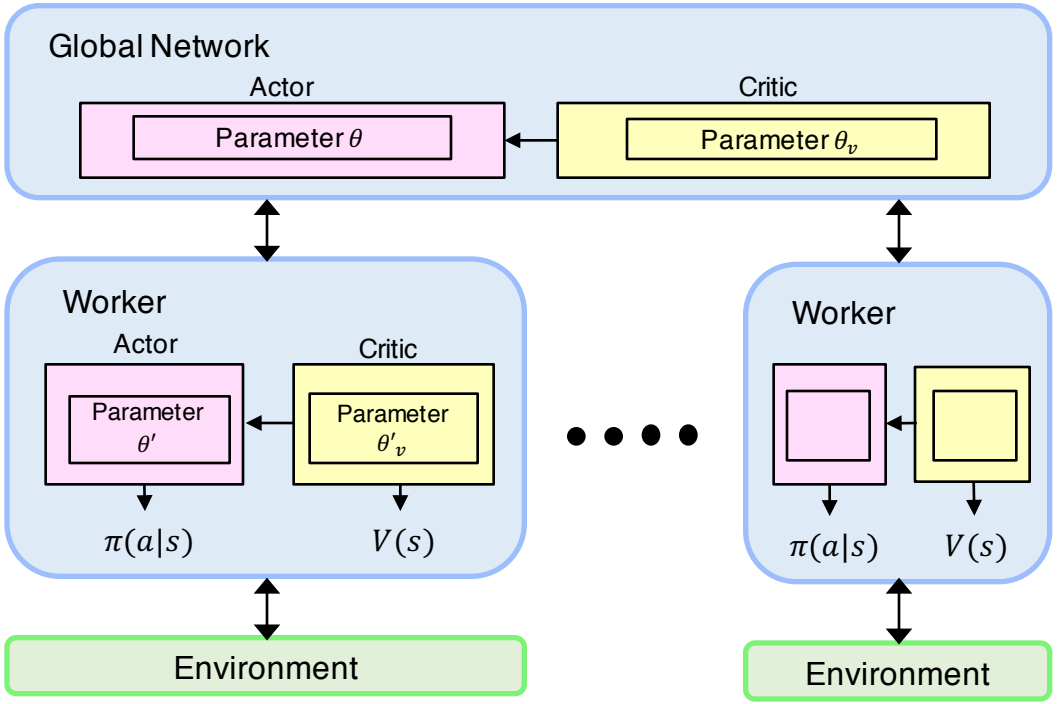
[Gu+, ICRA2016]

[Mnih+, Nature2015]

Asynchronous Advantage Actor-Critic [Mnih+, 2016]

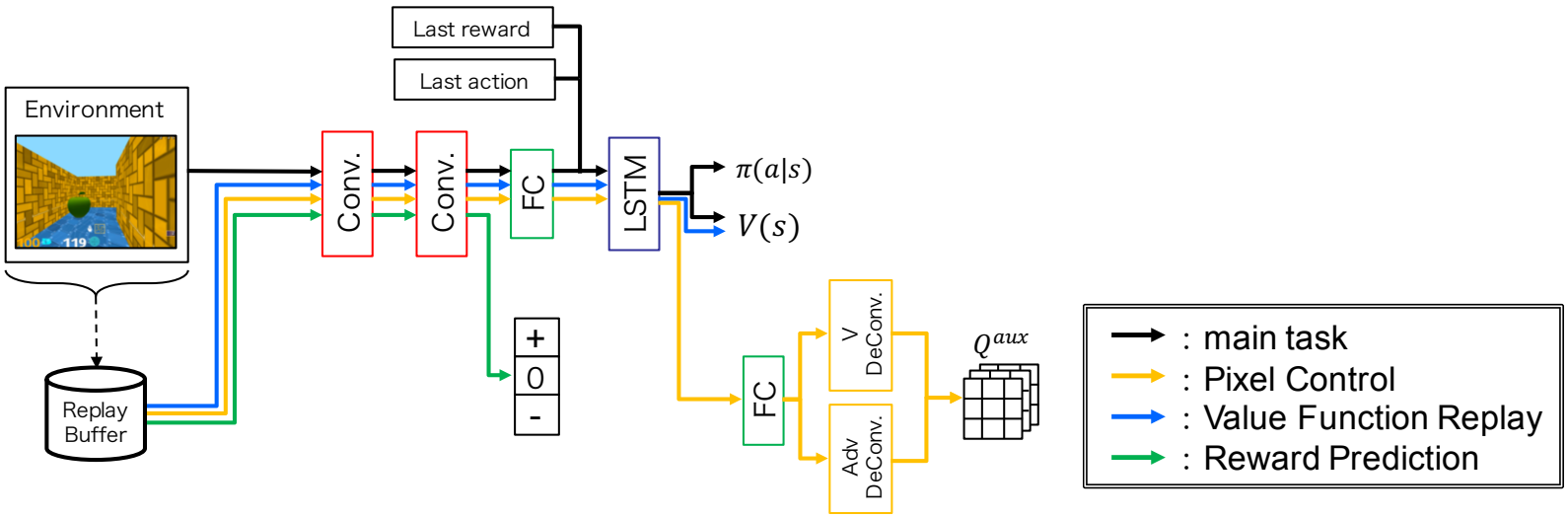
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

- Asynchronous
  - Each workers updates parameters asynchronously
- Advantage
  - Target error is calculate considering the reward more than 2 steps ahead in each worker
- Actor-Critic
  - Estimate
    - policy
    - State-value function



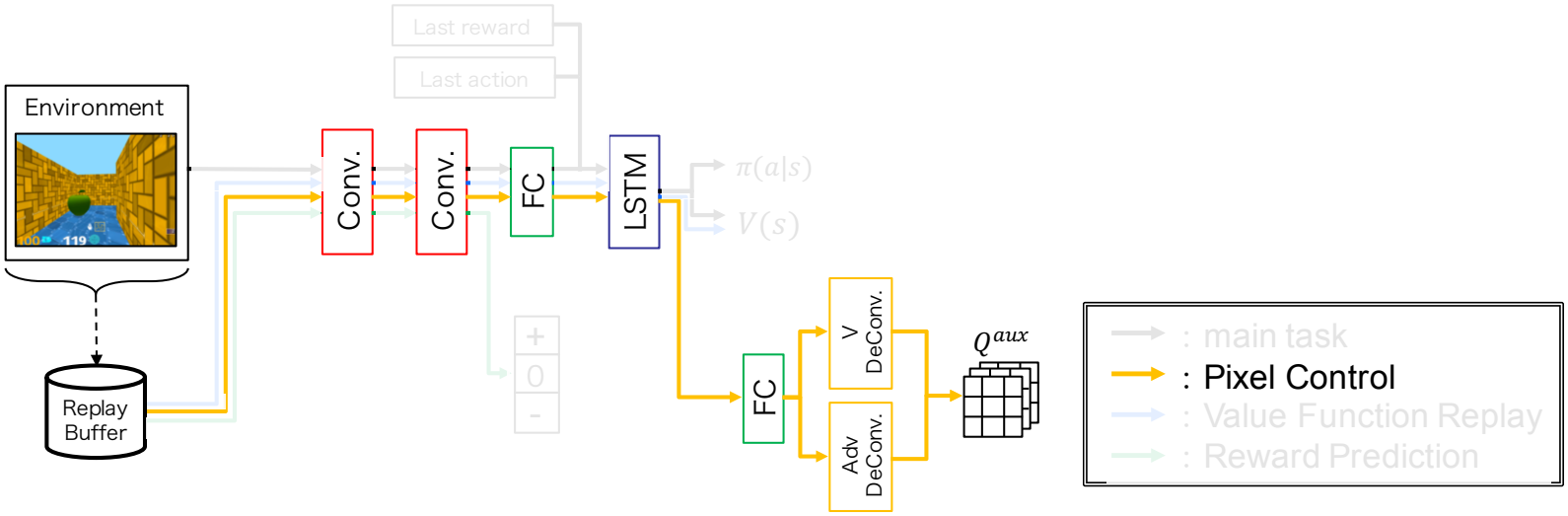
UNREAL [Jaderberg+, ICLR2017]

- Introducing three auxiliary tasks into the A3C
  - Pixel Control
    - Train actions that large changes in pixel values
  - Value Function Replay
    - Shuffle past experiences and train state-value functions
  - Reward Prediction
    - Predict future rewards



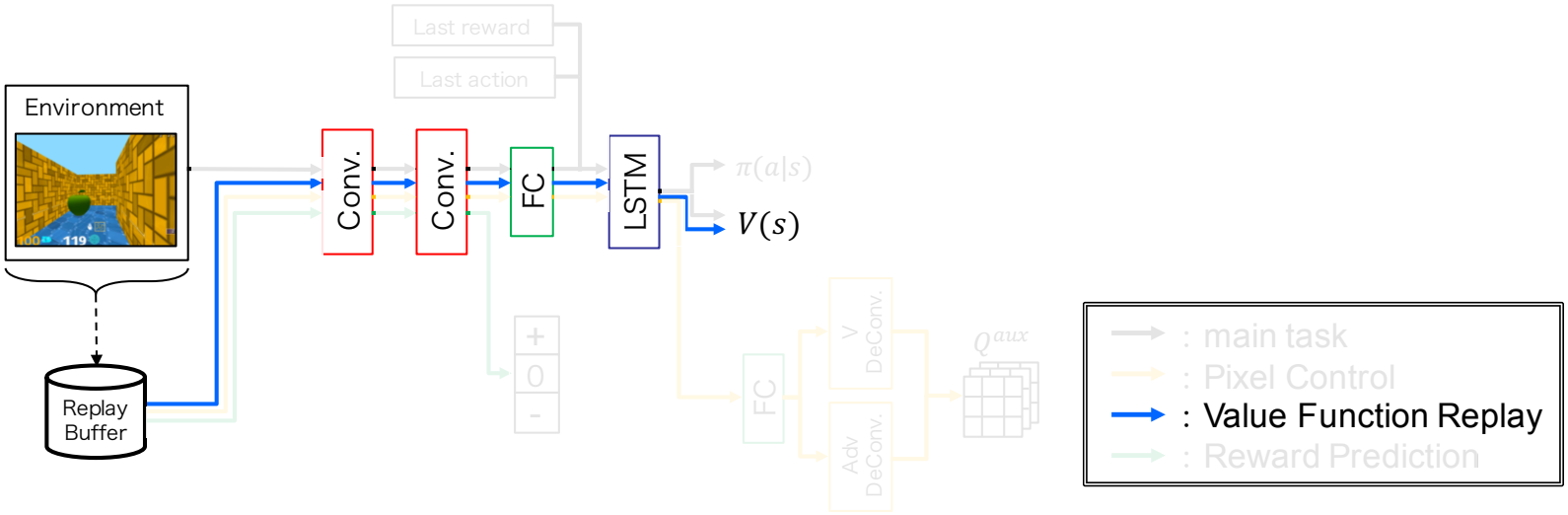
UNREAL [Jaderberg+, ICLR2017]

- Introducing three auxiliary tasks into the A3C
  - Pixel Control
    - Train actions that large changes in pixel values
  - Value Function Replay
    - Shuffle past experiences and train state-value functions
  - Reward Prediction
    - Predict future rewards



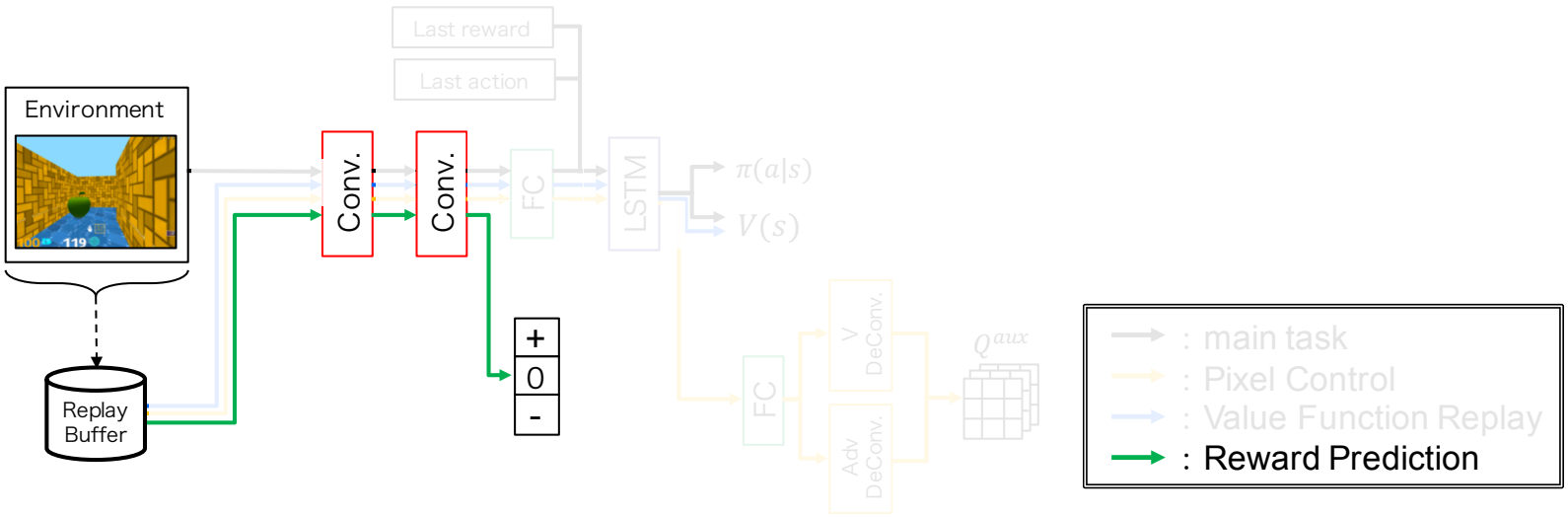
UNREAL [Jaderberg+, ICLR2017]

- Introducing three auxiliary tasks into the A3C
  - Pixel Control
    - Train actions that large changes in pixel values
  - Value Function Replay
    - Shuffle past experiences and train state-value functions
  - Reward Prediction
    - Predict future rewards



UNREAL [Jaderberg+, ICLR2017]

- Introducing three auxiliary tasks into the A3C
  - Pixel Control
    - Train actions that large changes in pixel values
  - Value Function Replay
    - Shuffle past experiences and train state-value functions
  - Reward Prediction
    - Predict future rewards



Loss function of UNREAL

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

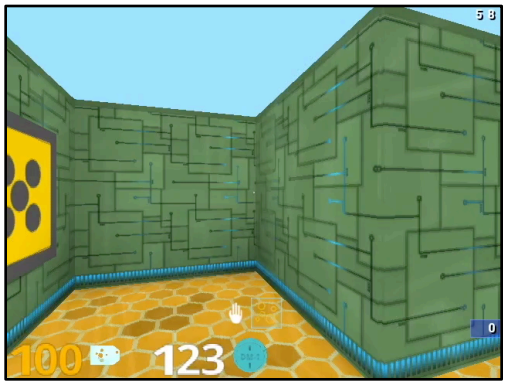
- The sum of main task loss and auxiliary tasks loss
  - L_{main} : Main task loss
  - $L_Q^{(c)}$: Pixel Control loss
  - L_{VR} : Value Function Replay loss
  - L_{RP} : Reward Prediction loss

$$L_{\text{UNREAL}} = \underbrace{L_{\text{main}}}_{\text{Main task}} + \underbrace{\sum_c L_Q^{(c)} + L_{\text{VR}} + L_{\text{RP}}}_{\text{Auxiliary Tasks}}$$

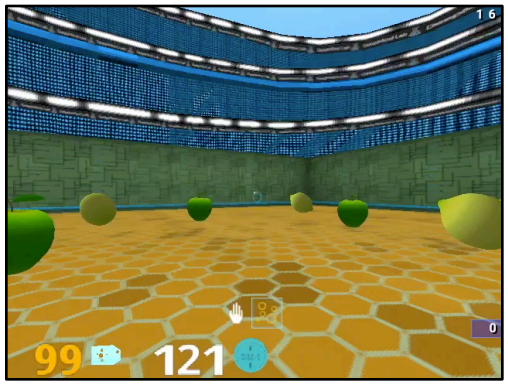
Preliminary experiment

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

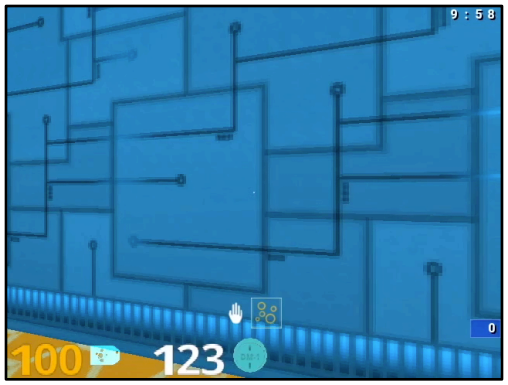
- Investigate whether each auxiliary task is effective or not
- Environment : DeepMind Lab [Beattie+, arXiv2016]
- Investigation functions
  - Pixel Control (PC)
  - Value Function Replay (VR)
  - Reward Prediction (RP)
  - Three auxiliary tasks (UNREAL)



nav_maze_static_01



seekavoid_arena_01



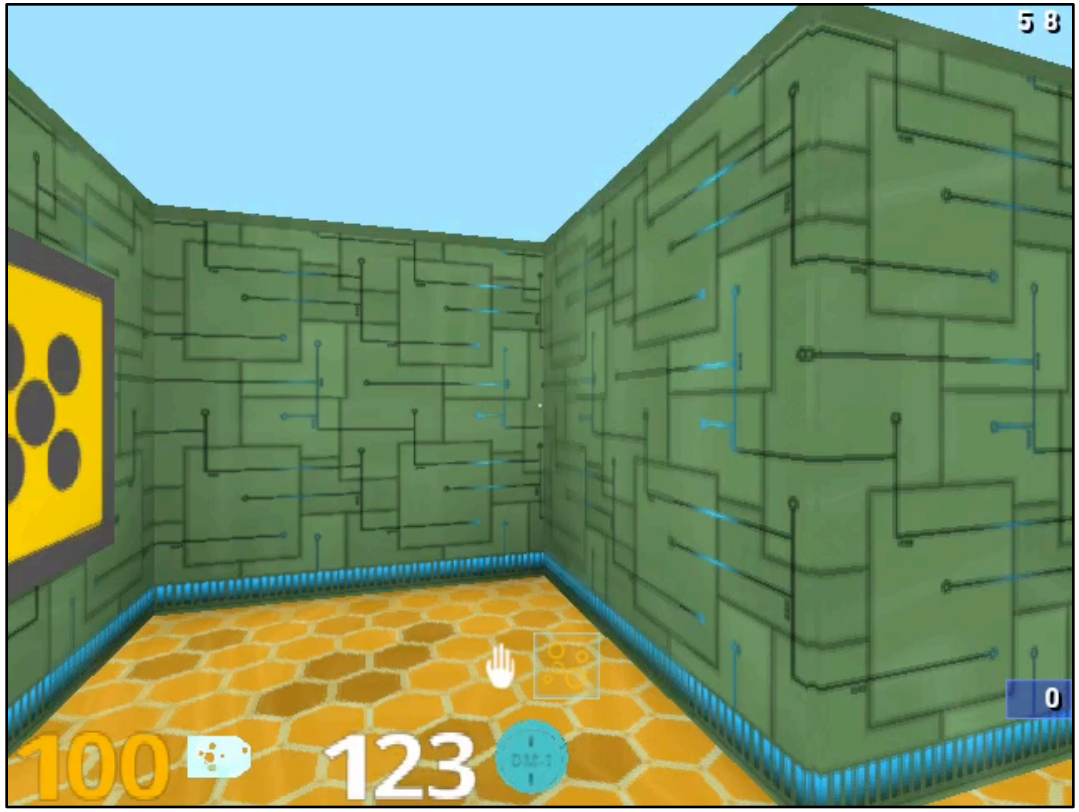
lt_horseshoe_color

nav_maze_static_01

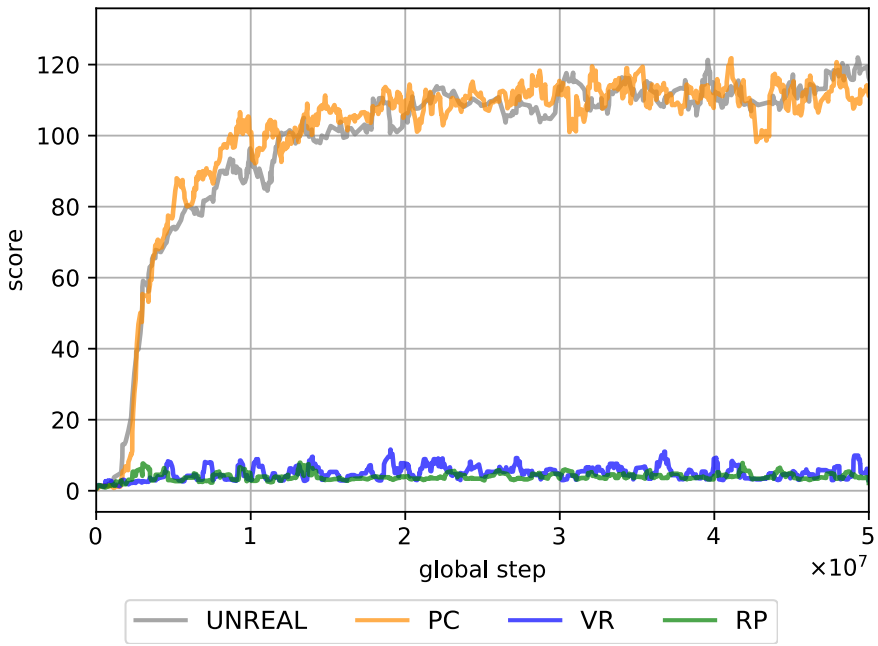
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

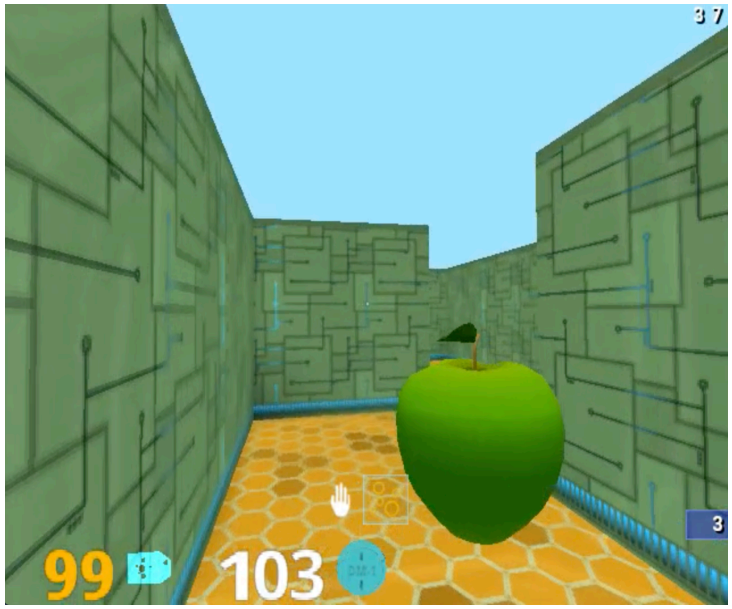
- A First-person viewpoint maze game

- Action
  - Look left
  - Look right
  - Forward
  - Backward
  - Strafe left
  - Strafe right
- Reward
  - Apple : +1
  - Goal : +10



Result (nav_maze_static_01)



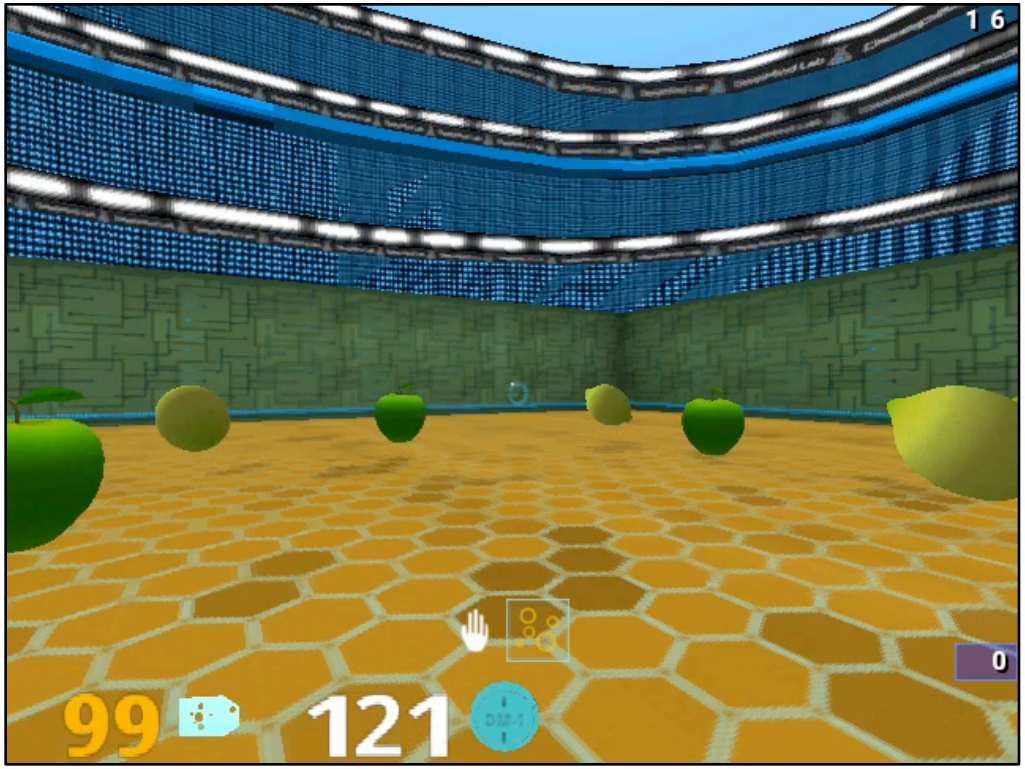


- Pixel Control is effective
  - Action changing pixel values promote movement

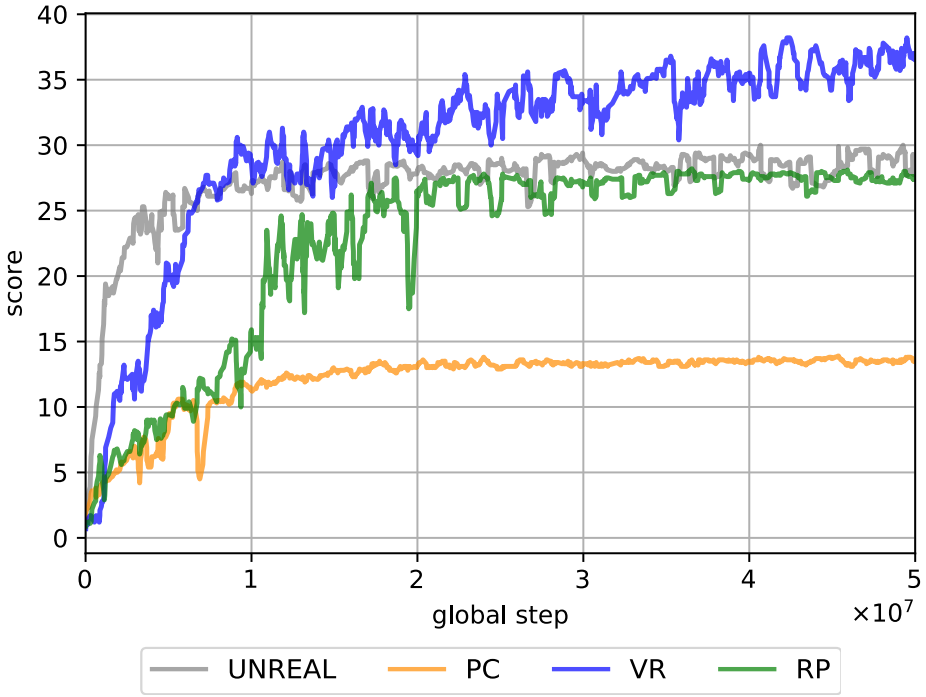
seekavoid_arena_01

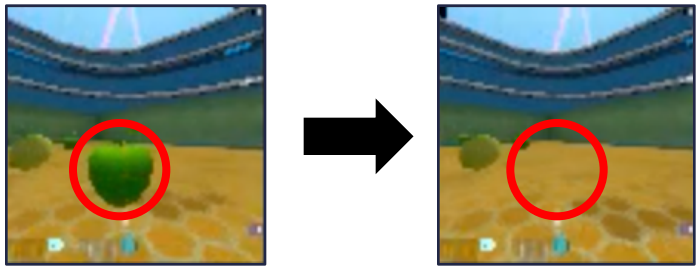
- Avoid lemons and earn apples game

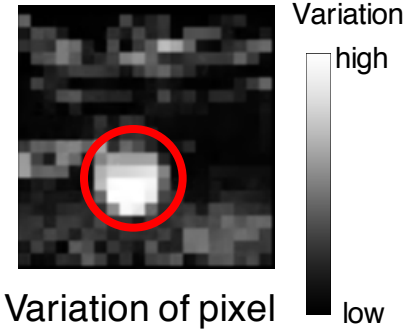
- Action
  - Look left
  - Look right
  - Forward
  - Backward
  - Strafe left
  - Strafe right
- Reward
  - Apple : +1
  - Lemon : -1



Result (seekavoid_arena_01)





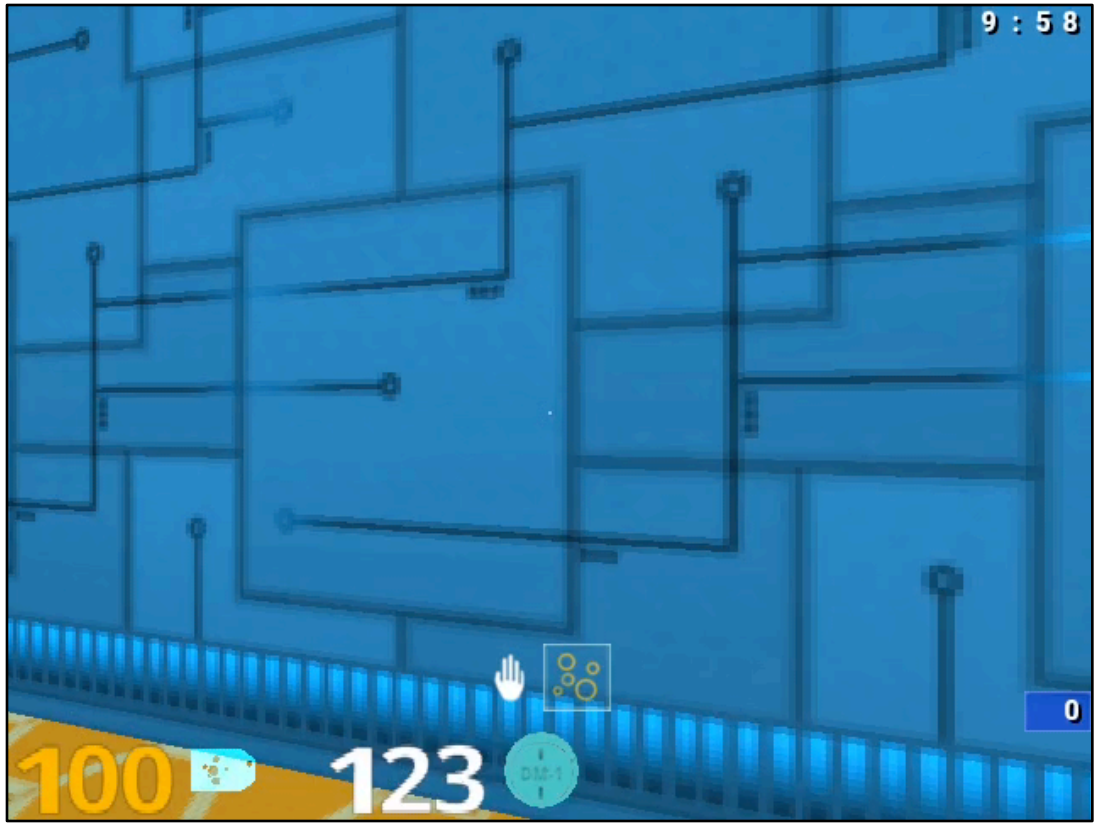


- Value Function Replay is effective
  - Actions changing pixel values are not suitable
  - Seekavoid obtains reward, frequently

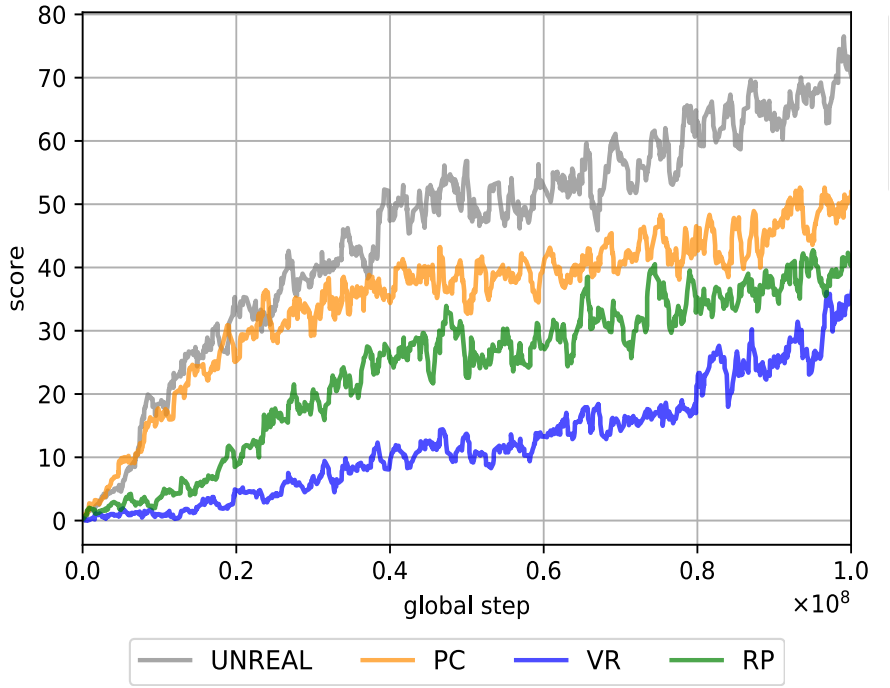
lt_horseshoe_color

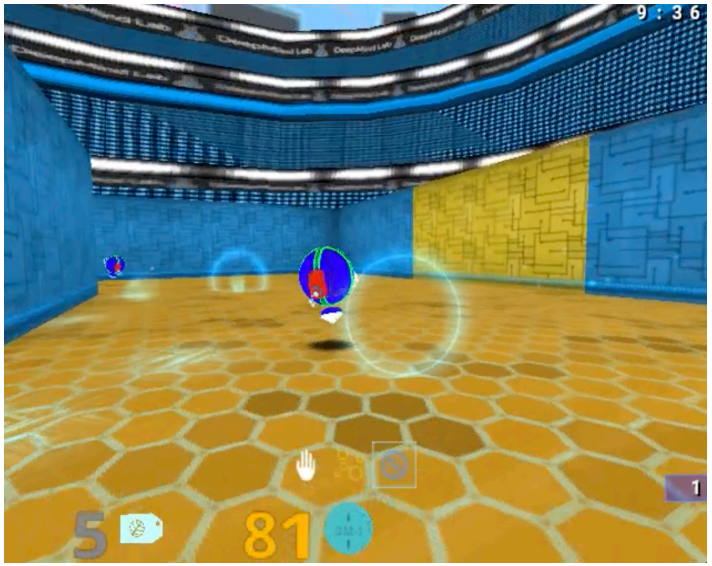
- First person shooting game

- Action
  - Look left
  - Look right
  - Forward
  - Backward
  - Strafe left
  - Strafe right
  - Attack
- Reward
  - Kill the enemy : +1



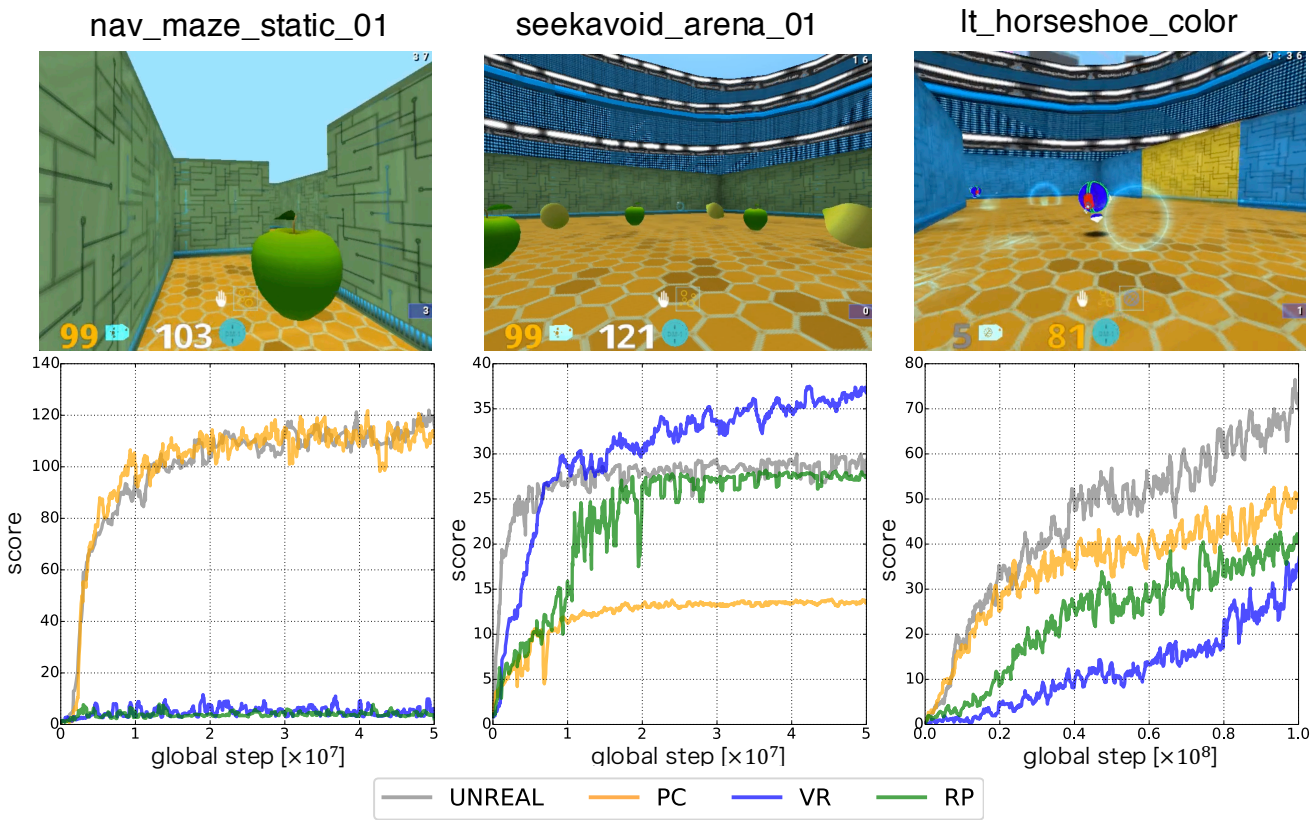
Result (lt_horseshoe_color)





- All auxiliary tasks are effective
  - Kill the enemy = Actions change pixel values
  - Reward (kill the enemy) acquired less frequent

Summary of pre-experiment

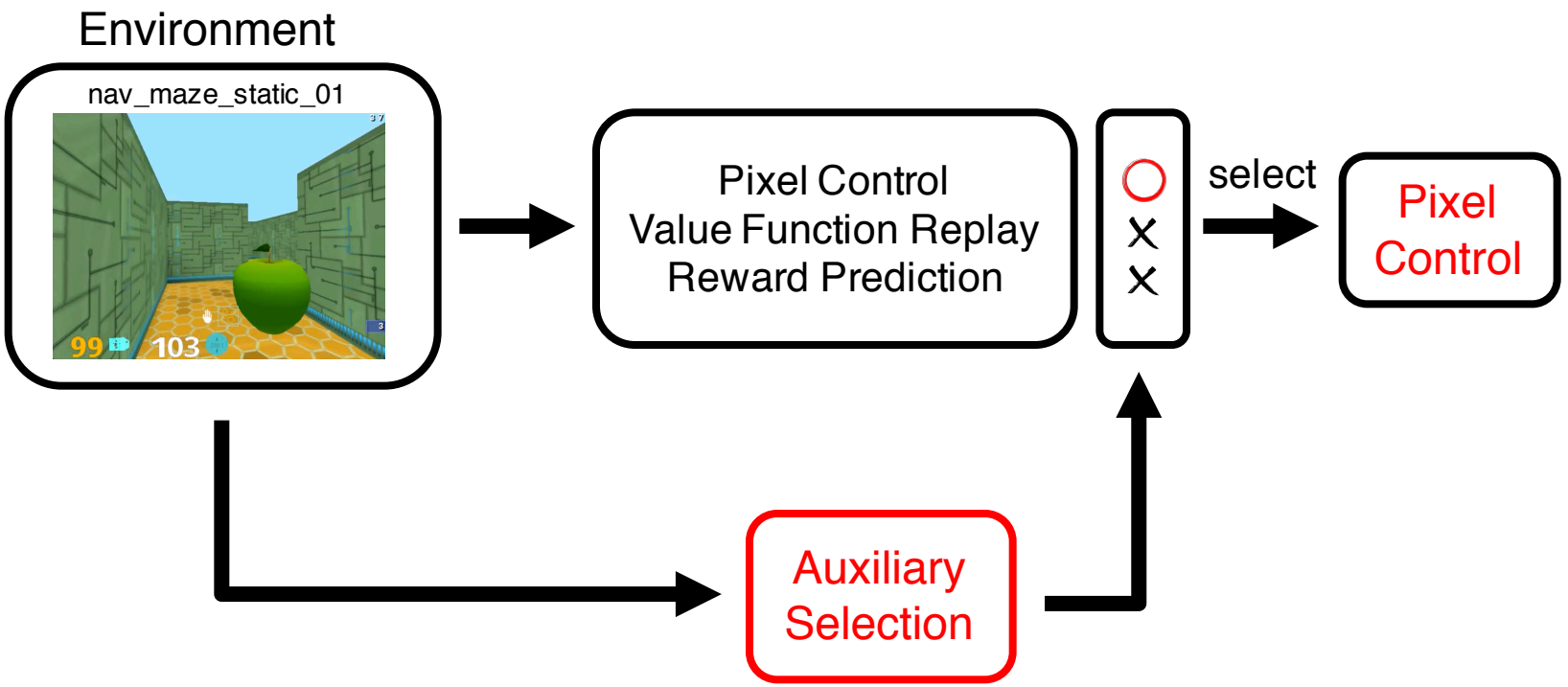


Optimal auxiliary task	Pixel Control UNREAL	Value Function Replay	UNREAL

→ Need to select suitable auxiliary tasks for game

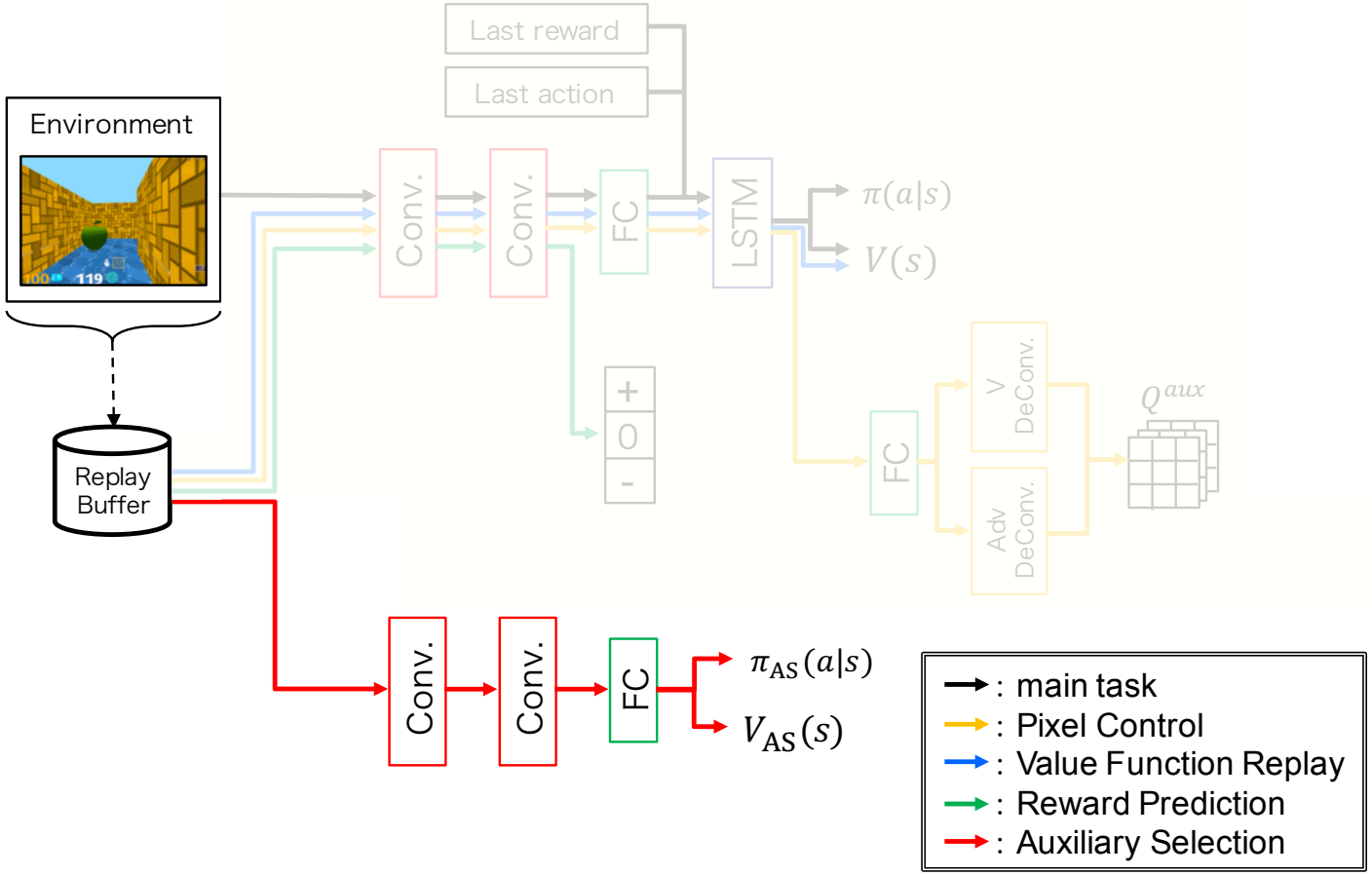
Purpose of proposed method

- Using only suitable auxiliary task for environment
  - Automatically select for suitable auxiliary tasks
- Proposed method
  - Auxiliary Selection
    - Adaptively selection of optimal auxiliary tasks



Auxiliary Selection

- A novel task to select the suitable auxiliary task for environment
  - Network build independent network from the main task



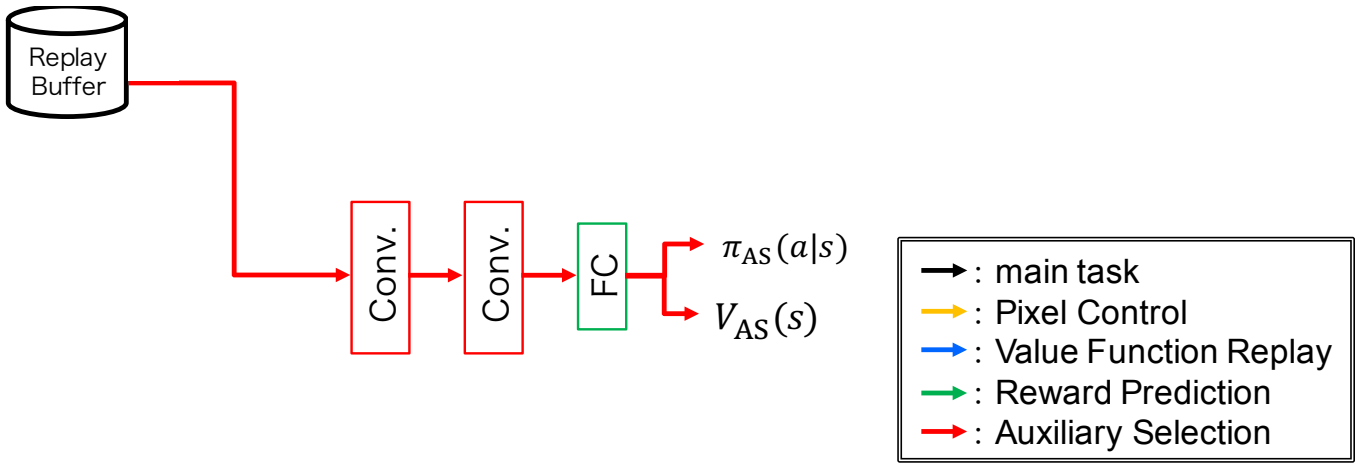
Action of Auxiliary Selection

- Weight of each auxiliary task $C_{\mathrm{PC}}, C_{\mathrm{VR}}, C_{\mathrm{RP}}$

$$(C_{\mathrm{PC}}, C_{\mathrm{VR}}, C_{\mathrm{RP}}) = (\{0,1\}, \{0,1\}, \{0,1\})$$

- Actions of Auxiliary Selection

$$\arg\max_{\pi_{\mathrm{AS}}} a = \{C_{\mathrm{PC}}, C_{\mathrm{VR}}, C_{\mathrm{RP}}\} = \underbrace{\{0,0,0\} \sim \{1,1,1\}}_{\text{8 patterns}}$$

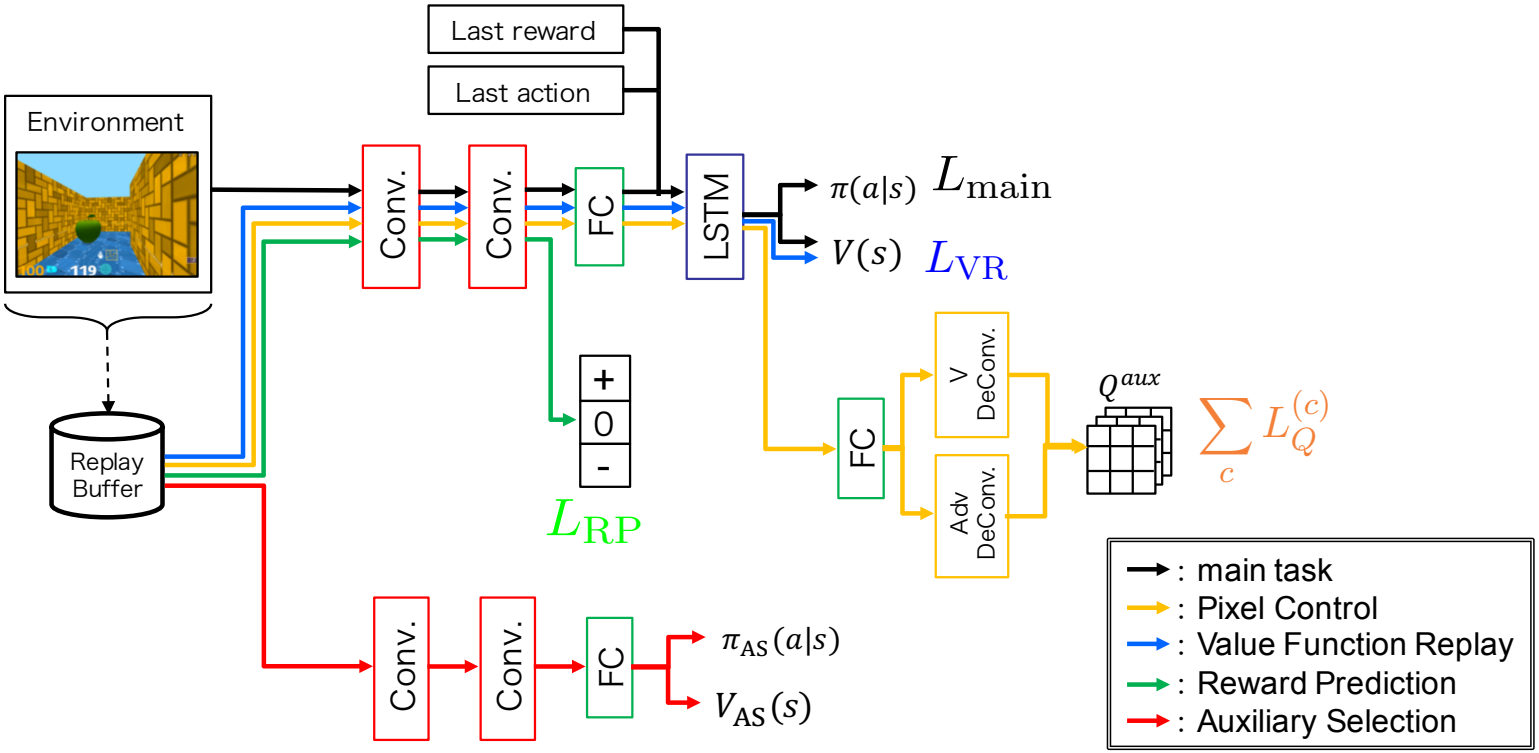


Loss of main and auxiliary tasks

- Multiply Auxiliary Selection outputs and loss of auxiliary tasks

$$L_{\mathrm{UNREAL}} = L_{\mathrm{main}} + C_{\mathrm{PC}} \sum_{c} L_Q^{(c)} + C_{\mathrm{VR}} L_{\mathrm{VR}} + C_{\mathrm{RP}} L_{\mathrm{RP}}$$

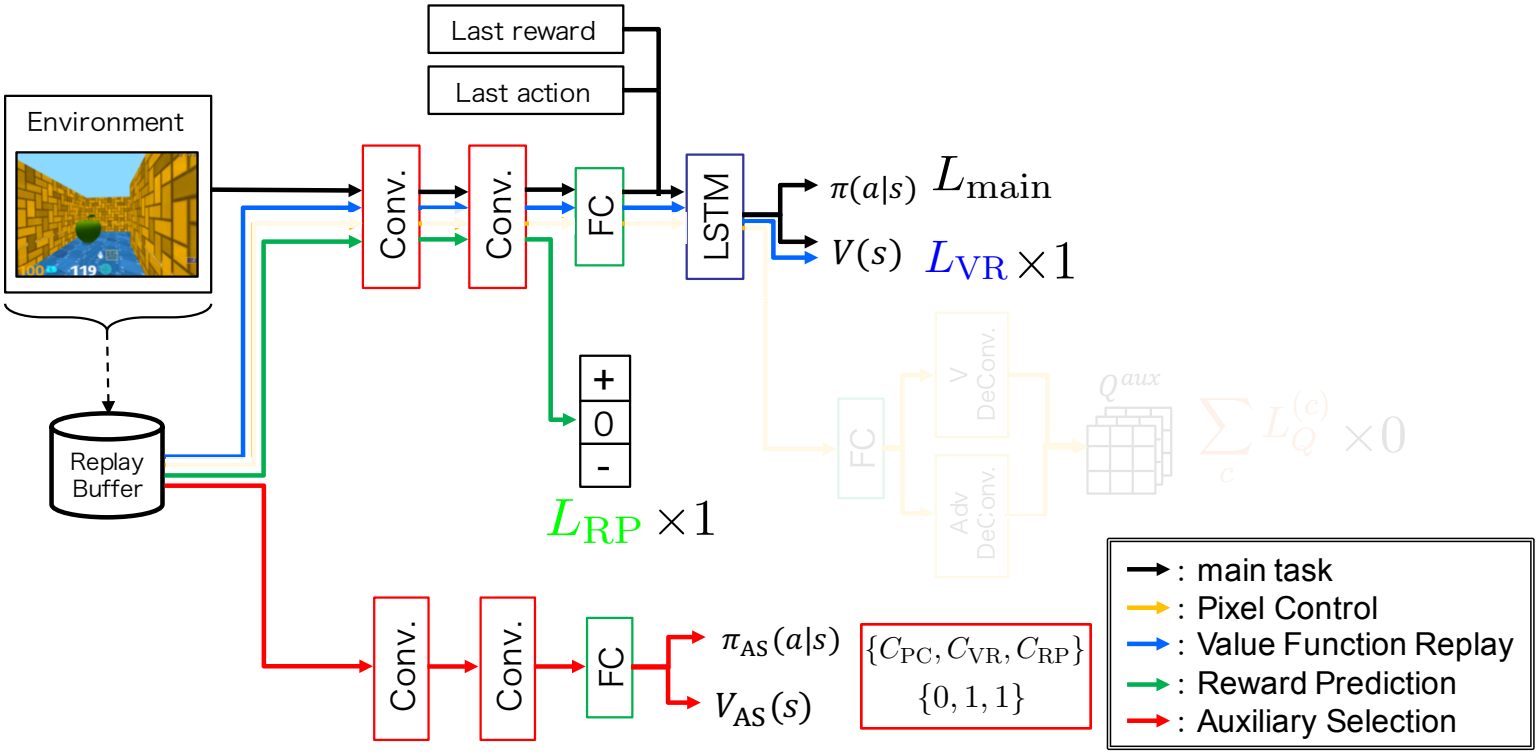
$\{0,1\}$ $\{0,1\}$ $\{0,1\}$



Loss of main and auxiliary tasks

- Multiply Auxiliary Selection outputs and loss of auxiliary tasks

$$L_{\mathrm{UNREAL}} = L_{\mathrm{main}} + \underset{\{0\}}{C_{\mathrm{PC}}} \sum_{c} L_Q^{(c)} + \underset{\{1\}}{C_{\mathrm{VR}}} L_{\mathrm{VR}} + \underset{\{1\}}{C_{\mathrm{RP}}} L_{\mathrm{RP}}$$

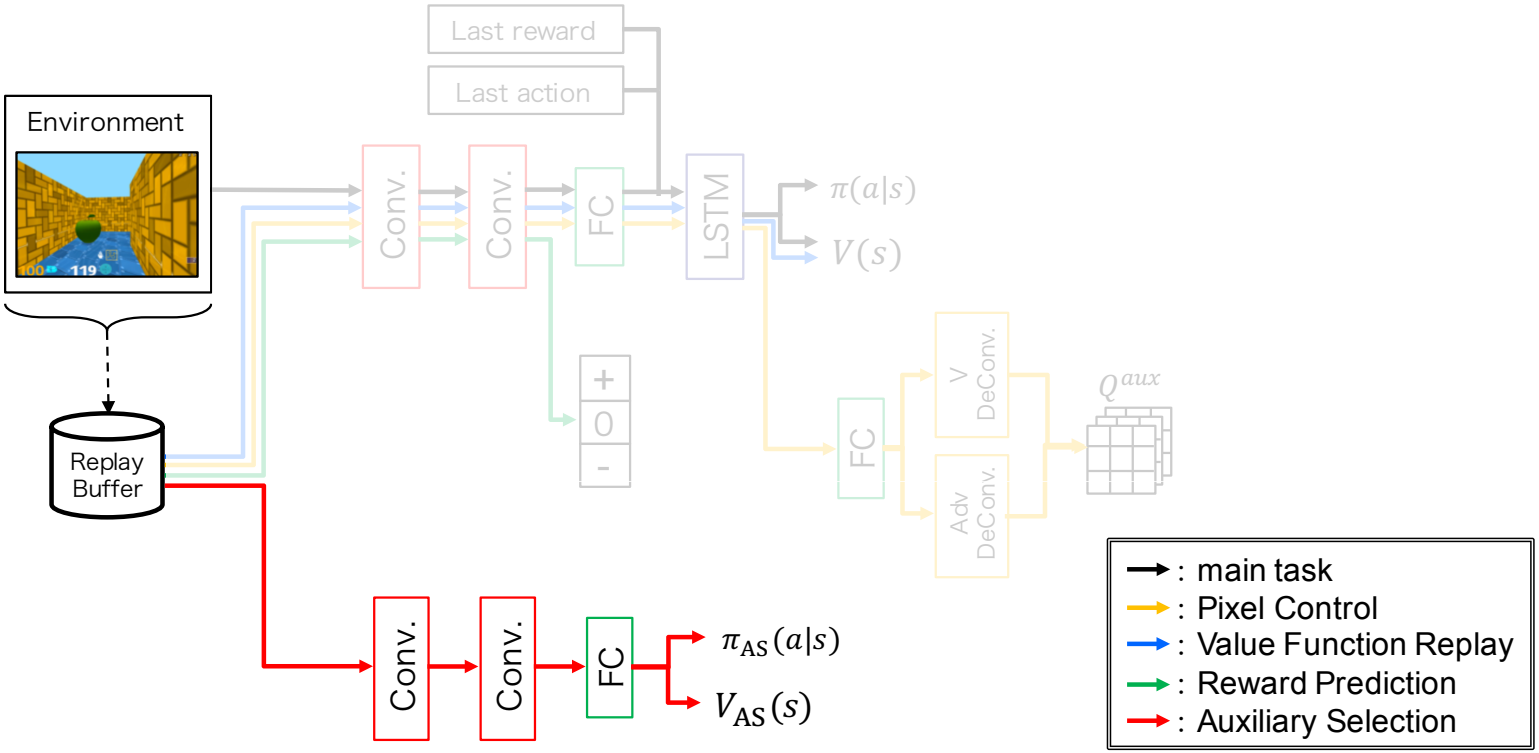


Loss function of Auxiliary Selection

- Adding losses of policy and state-value function

$$L_{\mathrm{AS}} = L(\pi_{\mathrm{AS}}) + L(V_{\mathrm{AS}})$$

Loss of policy　　Loss of state-value function

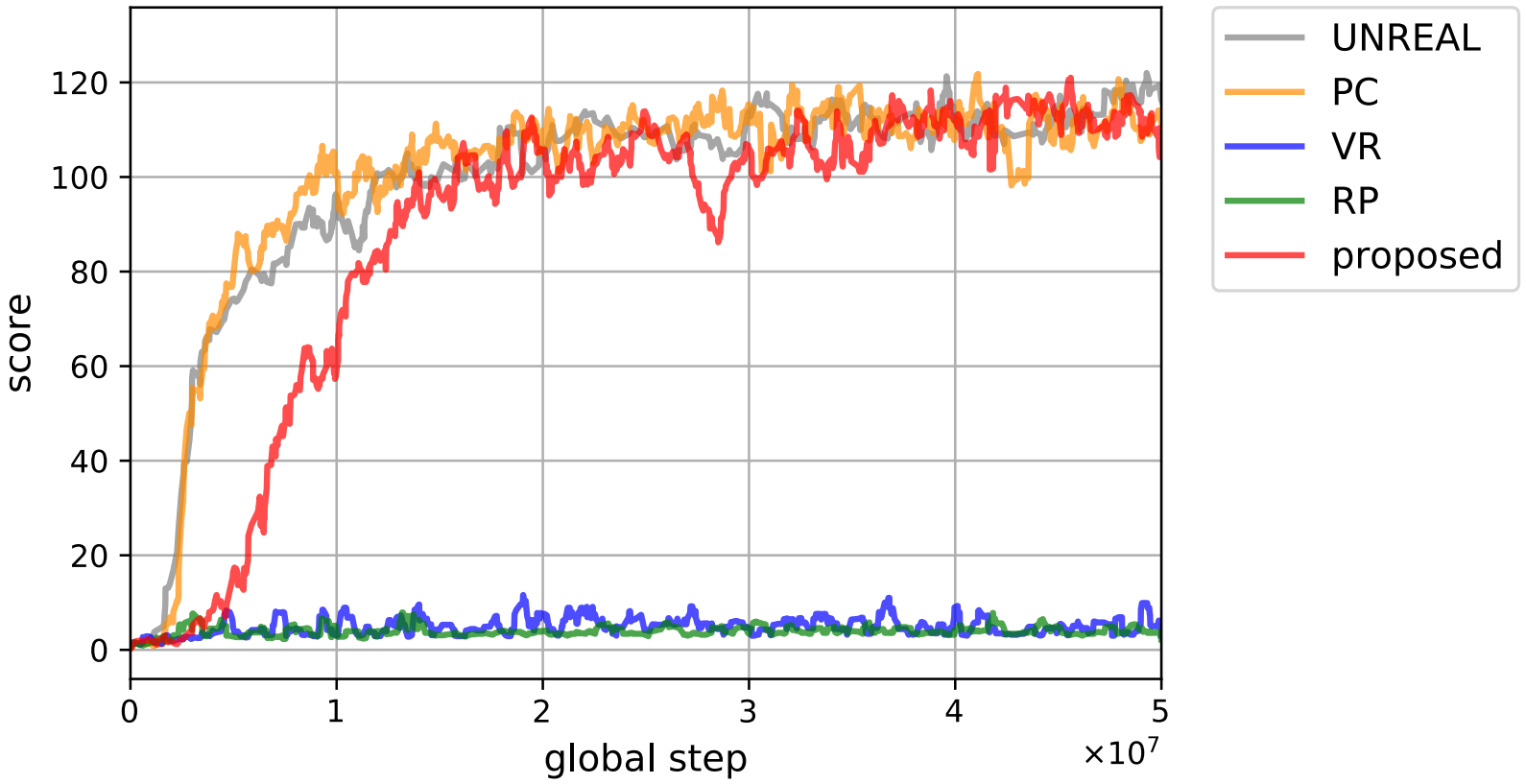


Experiment settings

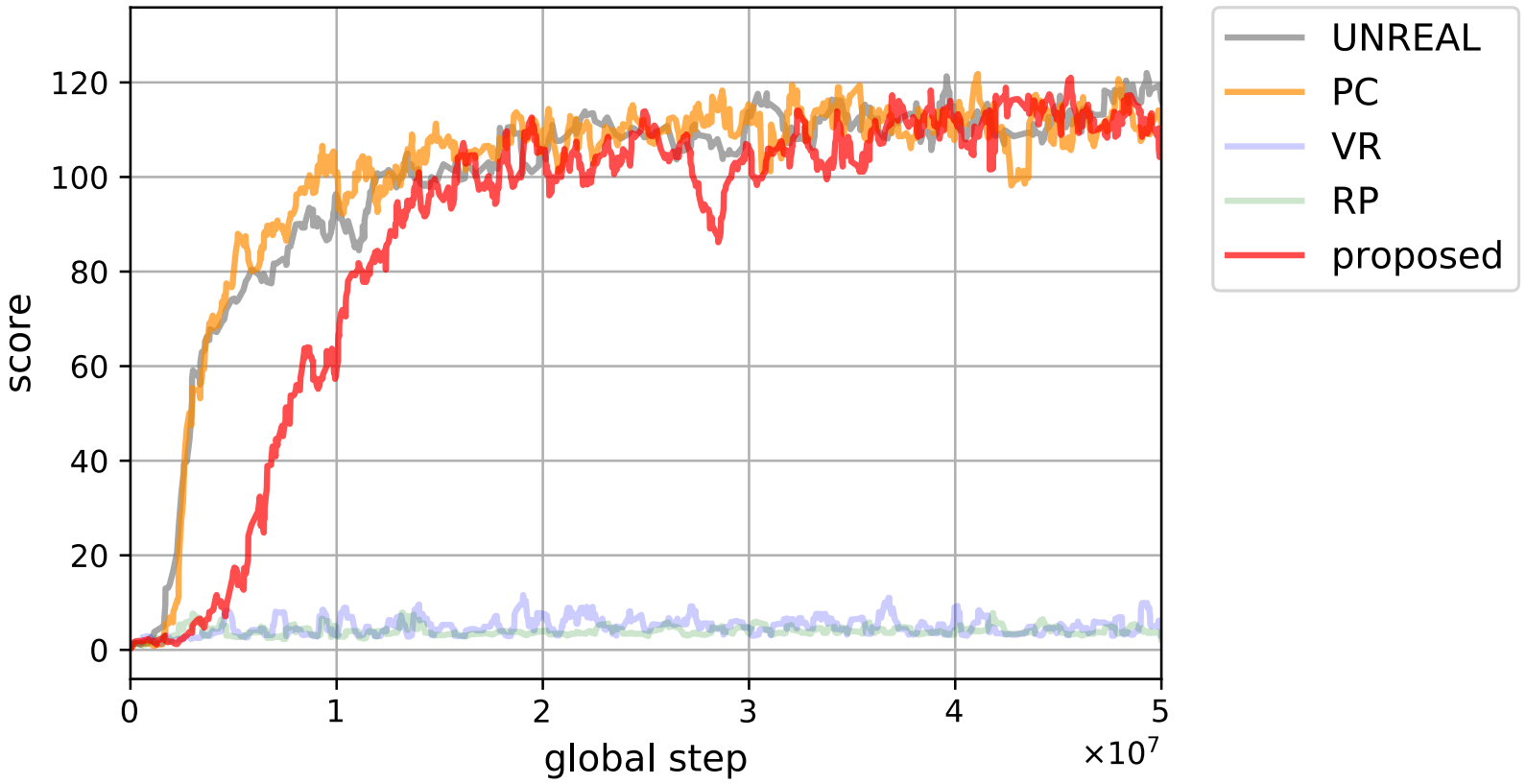
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

- Environment : DeepMind Lab [Beattie+, arXiv2016]
- Training setting
  - # of steps
    - 1.0×10^{7} steps (maze and seekavoid)
    - 1.0×10^{8} steps (horseshoe)
  - # of workers
    - 8
- Comparison
  - Only auxiliary task (PC, VR, RP)
  - Three auxiliary tasks (UNREAL)
  - Proposed method (proposed)

Result (nav_maze_static_01)

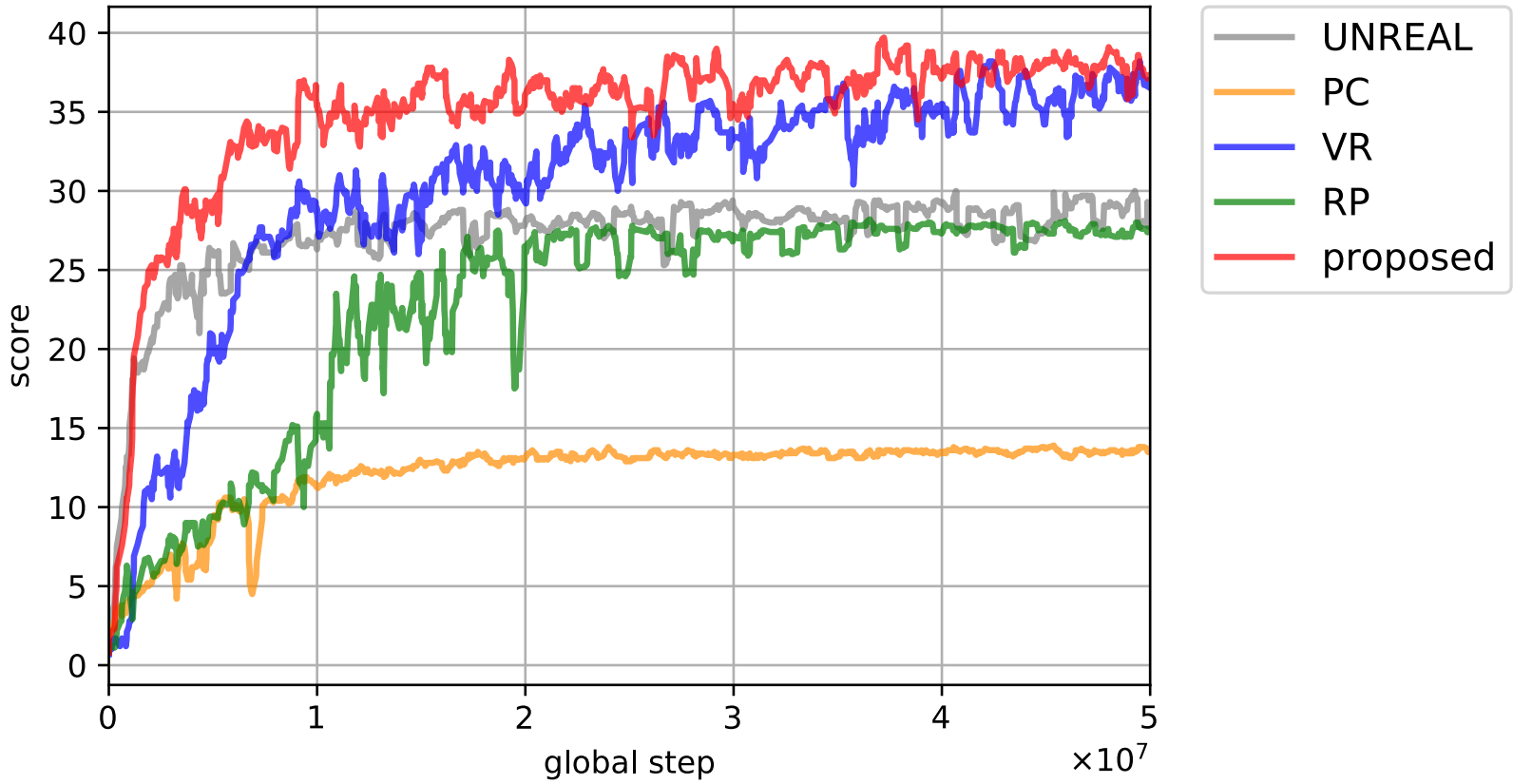


Result (nav_maze_static_01)

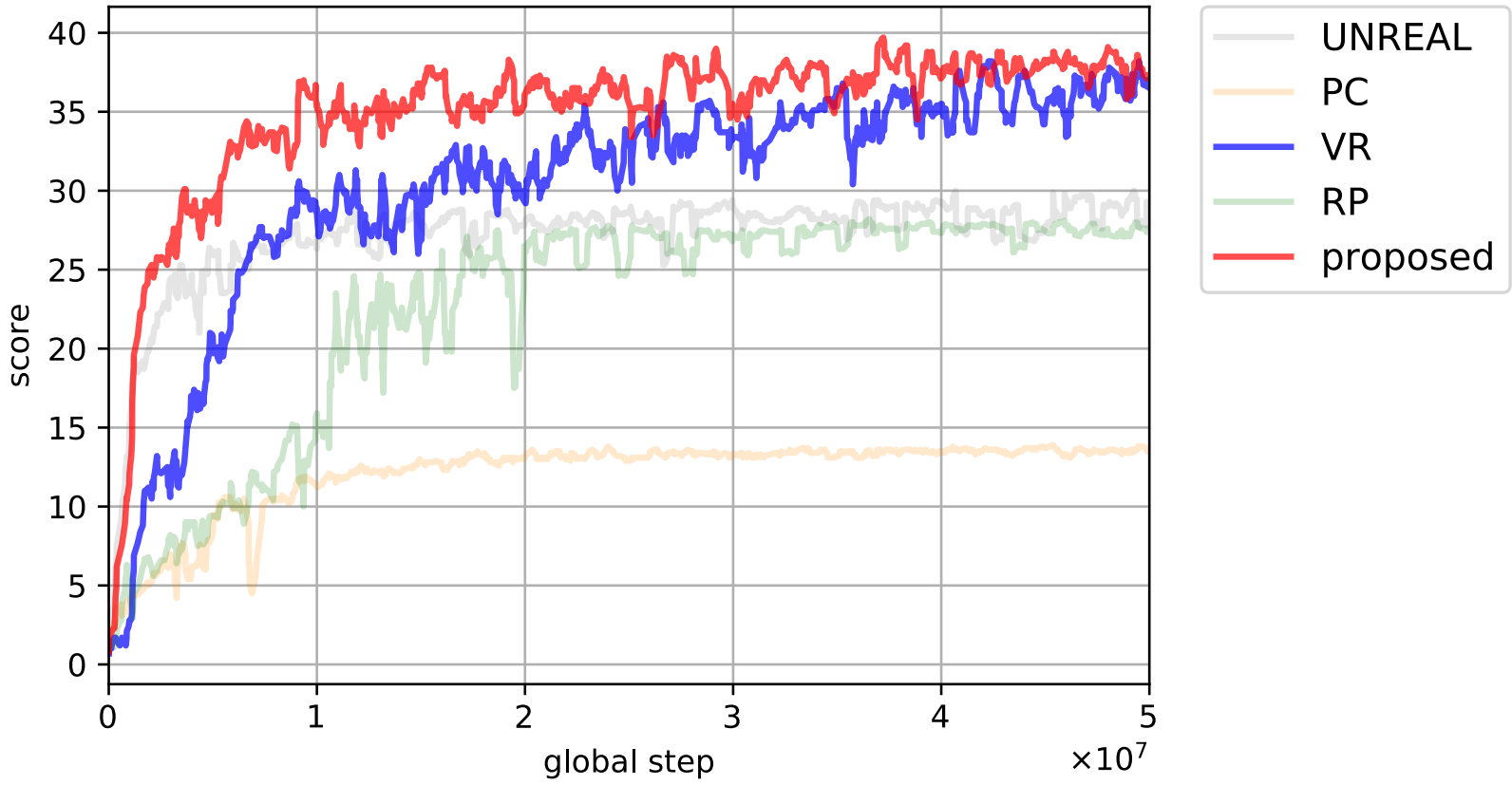


→ Proposed method achieve high score as same as UNREAL or PC

Result (seekavoid_arena_01)

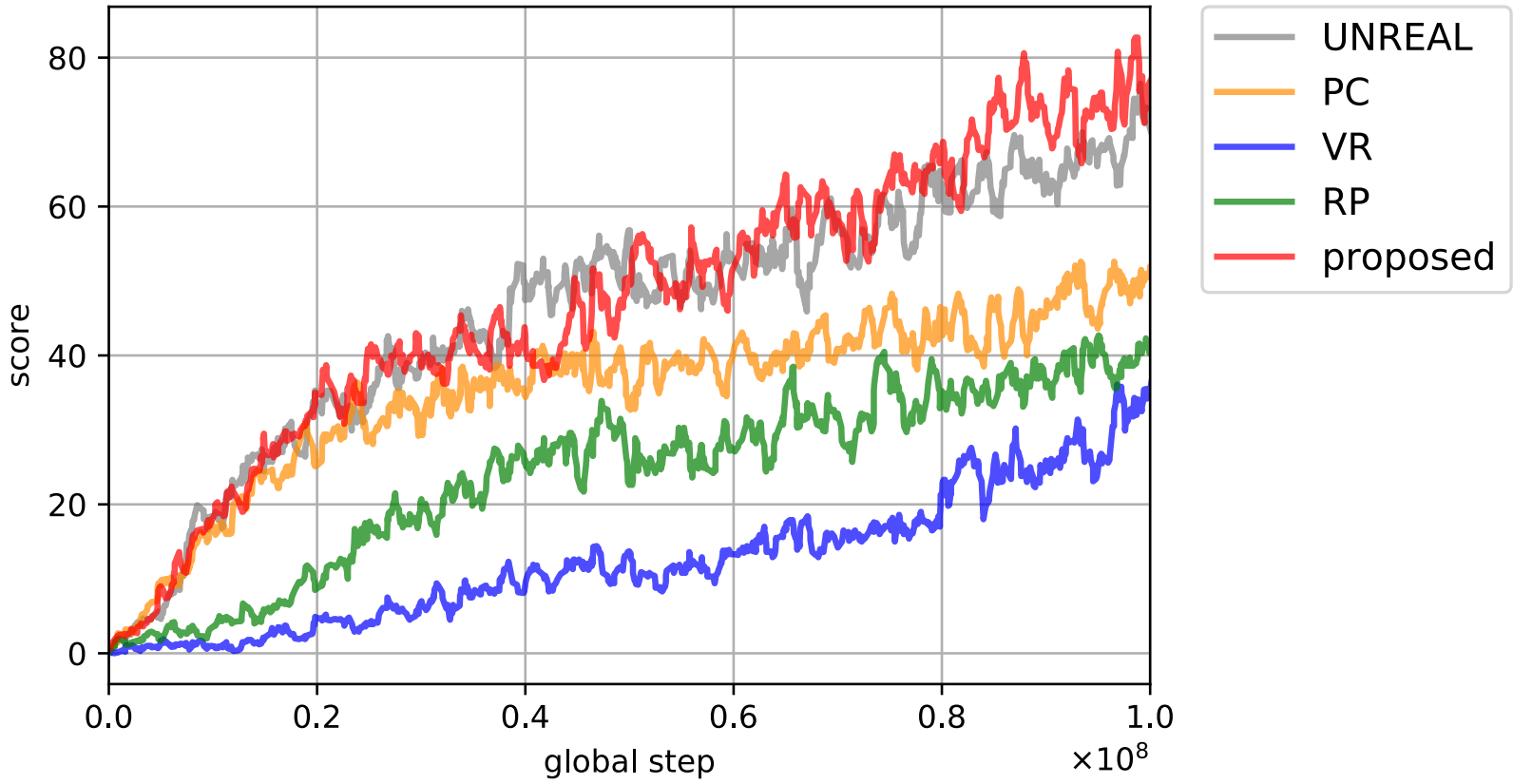


Result (seekavoid_arena_01)

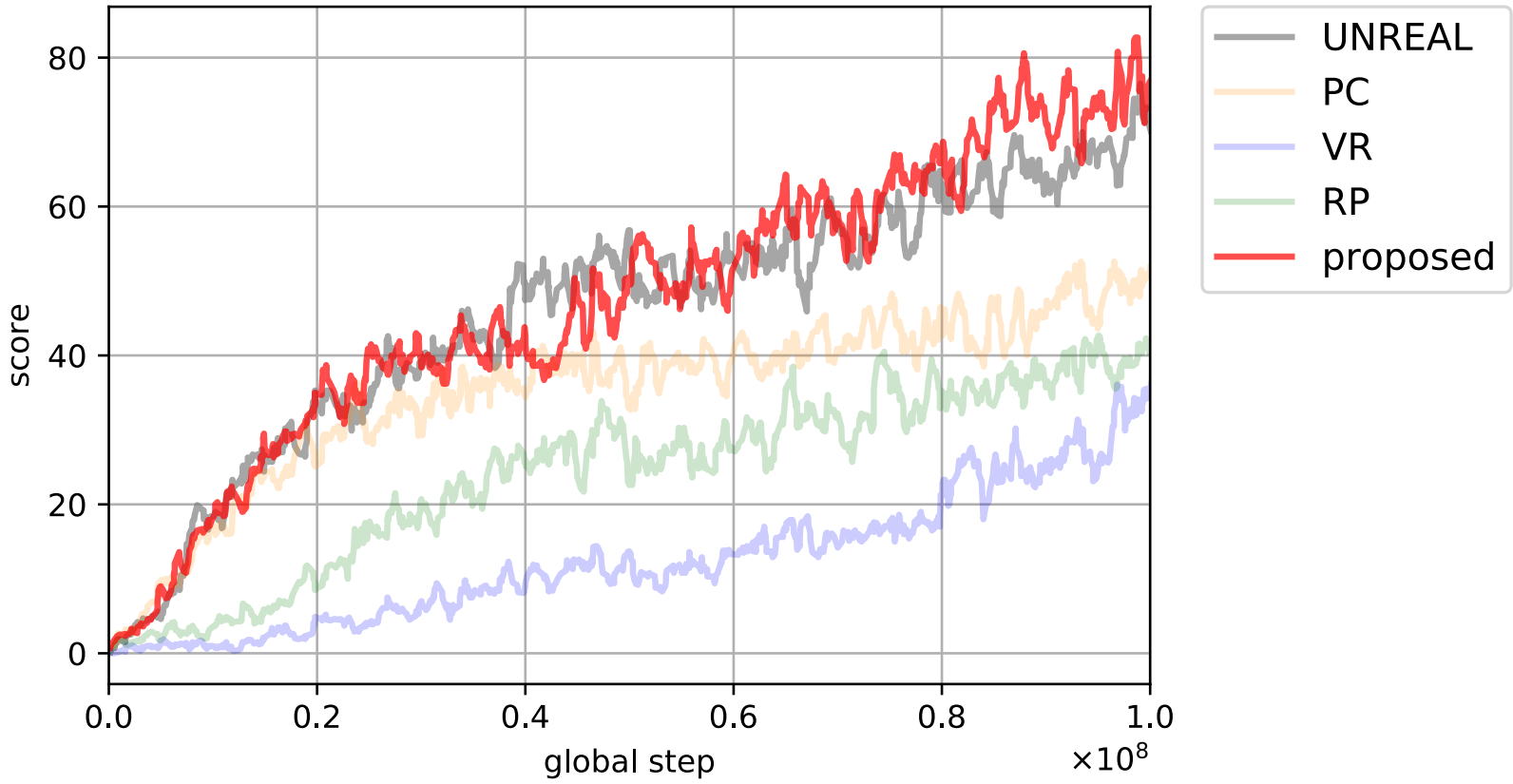


→ Proposed method achieve high score as same as VR

Result (lt_horseshoe_color)



Result (lt_horseshoe_color)



→ Proposed method achieve high score as same as UNREAL

Analysis of the selected auxiliary tasks (nav_maze_static_01)

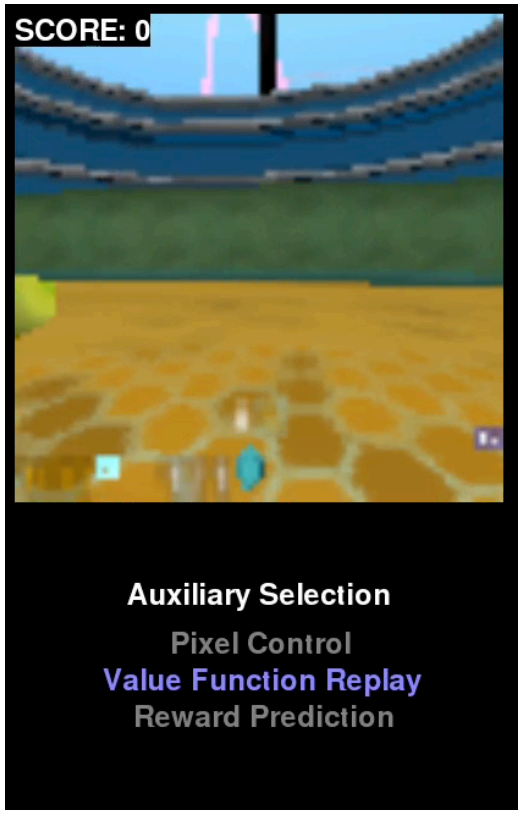
Selection percentage of each AT in one episode [%]

	Pixel Control	Value Function Replay	Reward Prediction
maze	48.3	54.1	41.0
seekavoid	0.1	100.0	0.0
horseshoe	94.9	0.1	99.9

※ 50 episodes average

→ All auxiliary tasks are equivalently selected

Analysis of the selected auxiliary tasks (seekavoid_arena_01)



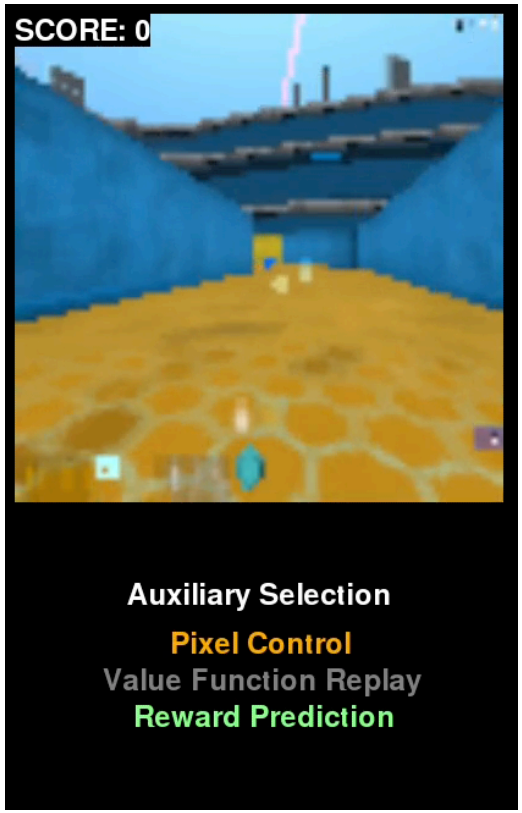
Selection percentage of each AT in one episode [%]

	Pixel Control	Value Function Replay	Reward Prediction
maze	48.3	54.1	41.0
seekavoid	0.1	100.0	0.0
horseshoe	94.9	0.1	99.9

※ 50 episodes average

→ VR is stably selected

Analysis of the selected auxiliary tasks (lt_horseshoe_color)



Selection percentage of each AT in one episode [%]

	Pixel Control	Value Function Replay	Reward Prediction
maze	48.3	54.1	41.0
seekavoid	0.1	100.0	0.0
horseshoe	94.9	0.1	99.9

※ 50 episodes average

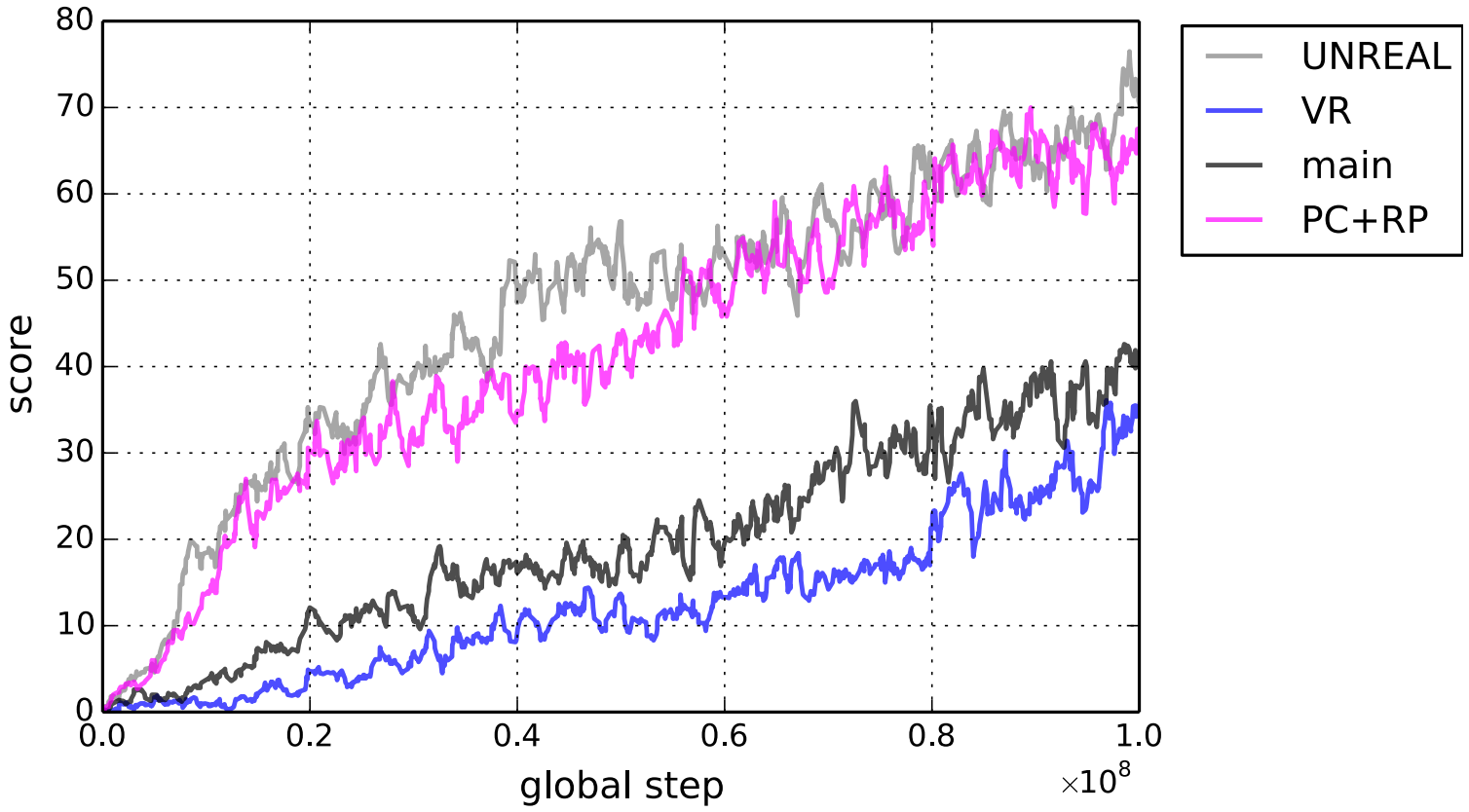
→ PC and RP are stably selected

Additional experiment

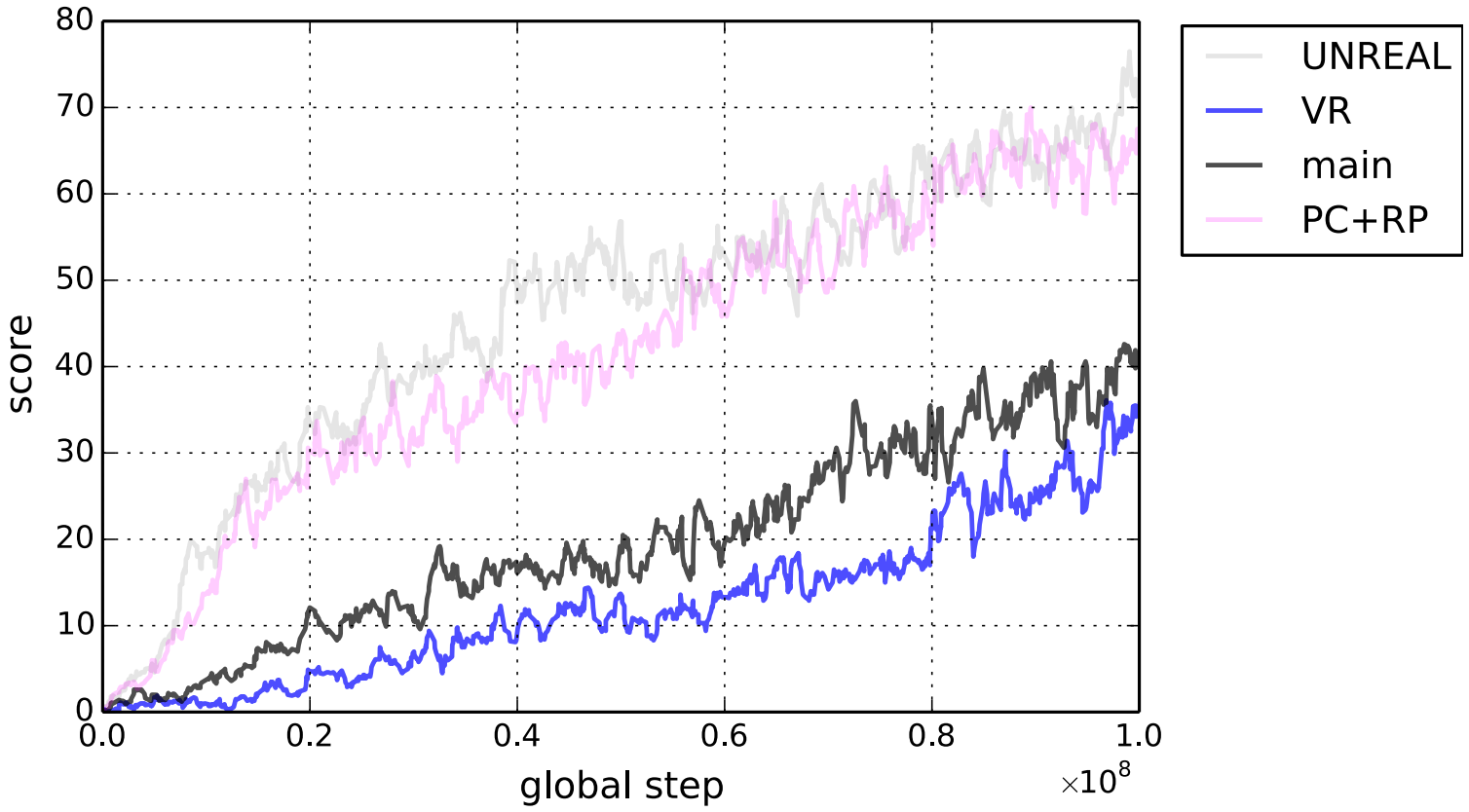
MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

- Investigation another combinations of auxiliary tasks
- Environment : lt_horseshoe_color (DeepMind Lab)
- Comparison : Compare scores in horseshoe
  - Three auxiliary tasks (UNREAL)
  - Value Function Replay (VR)
  - Pixel Control and Reward Prediction (PC+RP)
  - Only main task (main)

Result

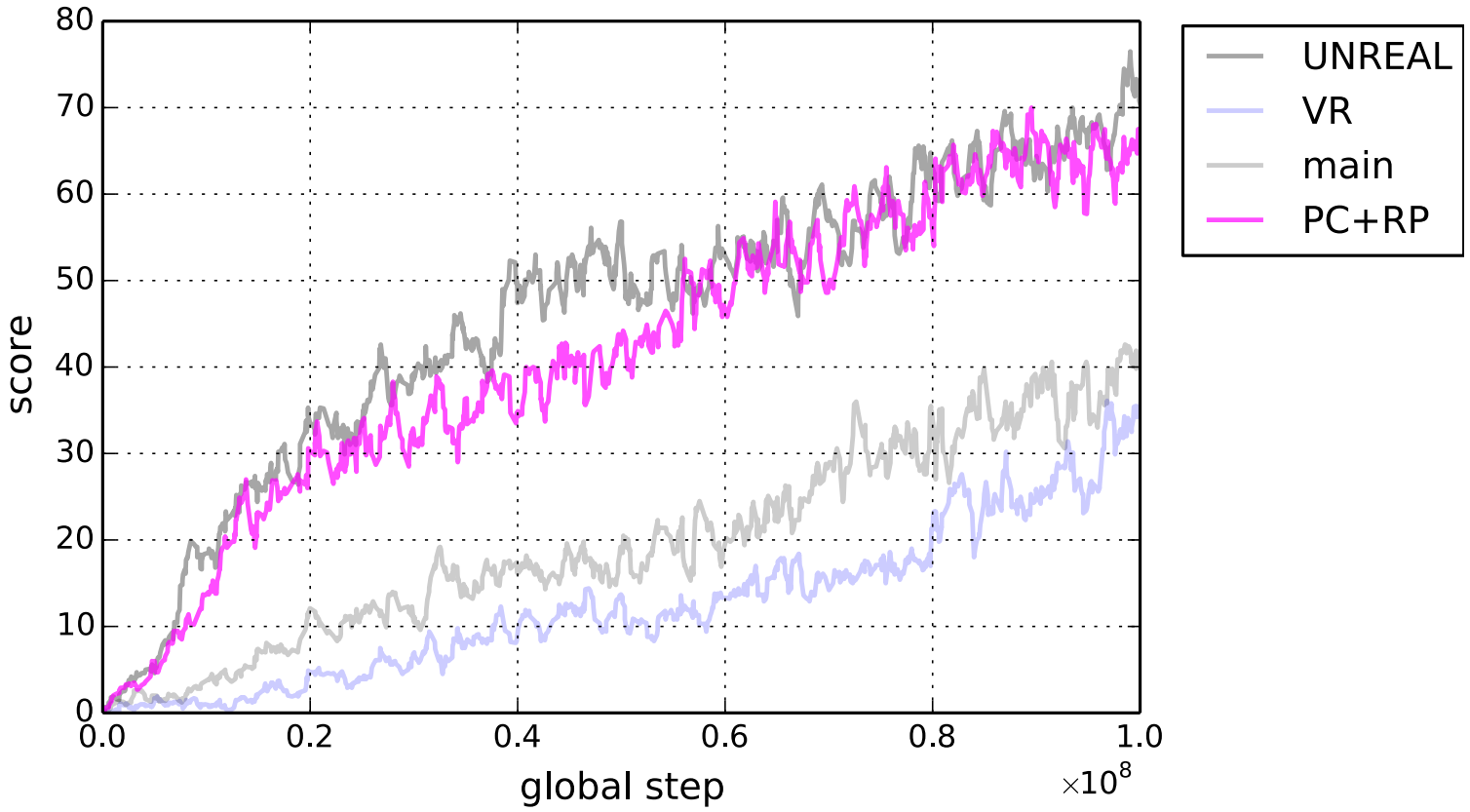

80
70
60
50
40
30
20
10
0
score
0.0
0.2
0.4
0.6
0.8
1.0
global step
×10^8
UNREAL
VR
main
PC+RP

Result



→ VR is lower than only main task

Result



→ VR is lower than only main task

→ PC+RP achieve high score as same as UNREAL

Conclusion

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

- Auxiliary Selection
  - Achieves the score as same as the optimal auxiliary task
  - Can select appropriate auxiliary tasks for each games
    - nav_maze_static_01 : UNREAL, Pixel Control
    - seekavoid_arena_01 : Value Function Replay
    - lt_horseshoe_color : Pixel Control + Reward Prediction

- Future work
  - Evaluating the proposed method in various environments with other auxiliary tasks