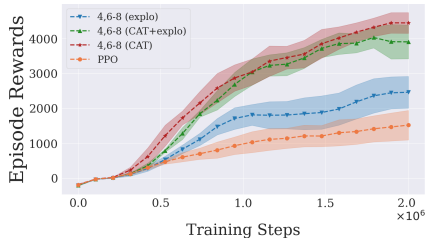
Cross-Domain Adaptive Transfer Reinforcement Learning Based on State-Action Correspondence (Supplementary material)

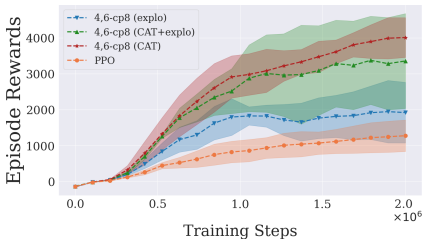
Heng You[1] **Tianpei Yang** *[1,2] **Yan Zheng**[1] **Jianye Hao** *[1] **Matthew E. Taylor**[2]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Department of Computing Science, University of Alberta and Alberta Machine Intelligence Institute, Canada

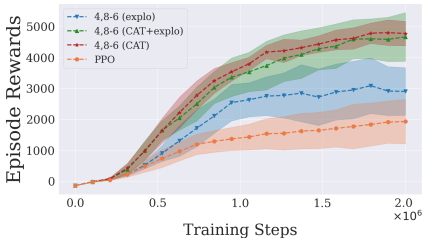
APPENDIX



(a) Target Env: CentipedeEight



(b) Target Env: CpCentipedeEight



(c) Target Env: CentipedeSix

Figure 1: Performance of different transfer manners including *explo*, CAT and their combination CAT + *explo*.

Experiments Description

For a Centipede agent, its state includes physical information such as joint angular velocity and twist angle, and its actions include control information for the torso bodies and legs, the same is true for the other two types of agents. See Table 1 for the state- action dimensions and source policy performance of all the agents, where "C-4" represents CentipedeFour. Although these agents have completely different state and action dimensions, they share the same dynamic principles, as well as similar physical structures and reward functions, which may be beneficial for transfer learning between different agents with different morphologies.

Table 1: The state-action dimensions and source policy performance of our environments.

Env	State Dim	Action Dim	Performance
C-4	97	10	2600
C-6	139	16	2000
C-8	181	22	1500
Cp-6	139	12	1610
Cp-8	181	18	1440
Ant	111	8	1100

Parameter Settings

The structure is the same for all networks: two fully-connected hidden layers both with 64 hidden units. See Table 2 for all the hyperparameters used in this paper.

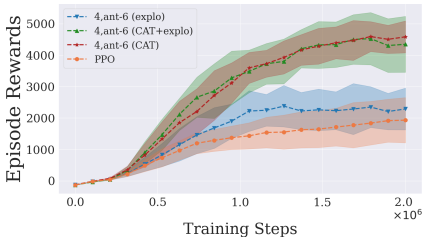
Table 2: CAT hyperparameters.

Hyperparameter	Value
Discount factor (γ)	0.99
Activation	tanh
Optimizer	Adam
Learning Rate	3 x 10^{-4}
Clip range (ε)	0.2
Evaluate steps	200
Batch size	64

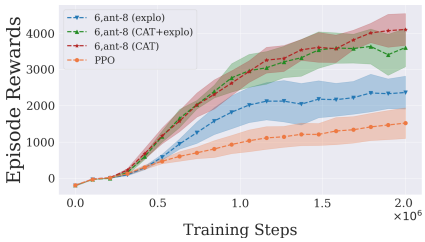
Different Transfer Manners

*Correspondence to Tianpei Yang <tpyang@tju.edu.cn>, Jianye Hao <jianye.hao@tju.edu.cn>

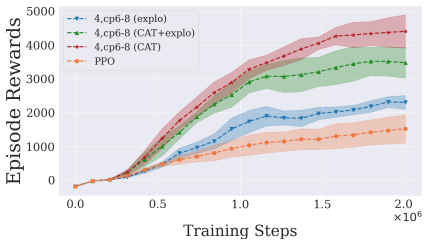
Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).



(a) Target Env: CentipedeSix



(b) Target Env: CentipedeEight



(c) Target Env: CentipedeEight

Figure 2: Performance of different transfer manners including *explo*, CAT and their combination CAT + *explo*.

In this experiment, we apply the two major transfer methods mentioned in Section 1 to the cross-domain setting through the learnt state-action correspondence to validate the choice in this paper. All the different transfer methods and their combination are as follows:

- *explo*: Reusing source policies to interact with the environment for exploration at a decreasing rate.
- CAT: Distilling knowledge from multiple source policy networks into the middle layers of the target policy networks used in CAT.
- CAT + *explo*: Applying *explo* while distilling knowledge from source policy networks.

Figure 1 and 2 show the performance of different transfer methods. We can see that the performance of *explo* is the worst except PPO since *explo* only selects one source policy at the same time to help the target task for exploration, which is an insufficient and ineffective method compared to CAT. In contrast, the CAT agent extracts useful knowledge of each source policy by combining knowledge from source policy networks through the adaptive weighting factors, thus outperforms all methods. Finally, we can see that the performance of CAT + *explo* is slightly lower than CAT in most cases. From our point of view, this is because *explo* reduces the effectiveness of CAT for the reasons we mentioned above. The results validate our choice in this paper.