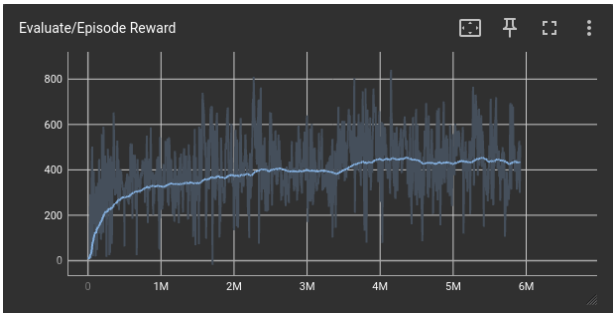
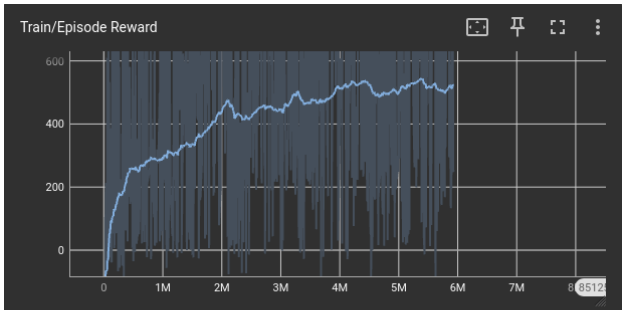
RL_LAB_4_313552041_洪日昇

Experimental Results (30%)

(1) Screenshot of Tensorboard training curve and testing results on TD3.

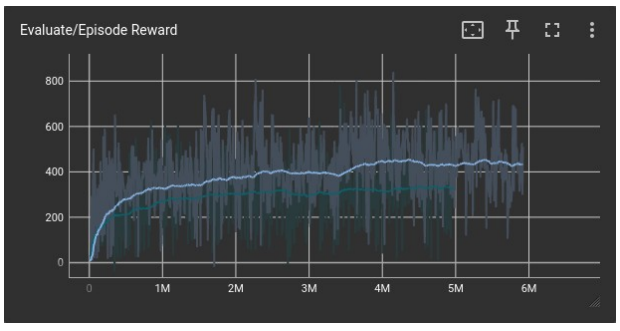


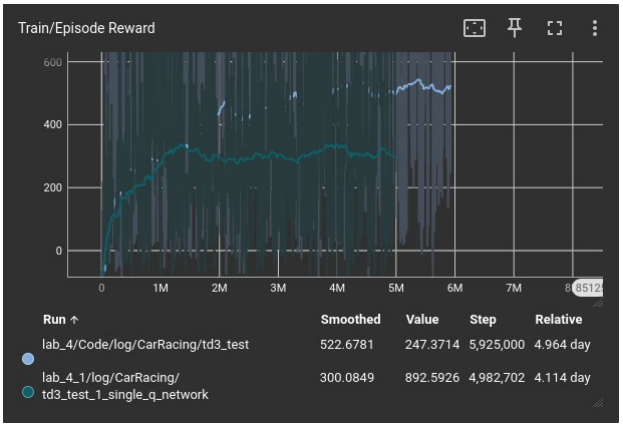



```
Evaluating...
Episode: 1      Length: 999     Total reward: 861.94
Episode: 2      Length: 513     Total reward: 604.65
Episode: 3      Length: 730     Total reward: 926.90
Episode: 4      Length: 605     Total reward: 939.40
Episode: 5      Length: 676     Total reward: 932.30
Episode: 6      Length: 269     Total reward: 252.28
Episode: 7      Length: 999     Total reward: 890.54
Episode: 8      Length: 688     Total reward: 931.10
Episode: 9      Length: 999     Total reward: 882.14
Episode: 10     Length: 999     Total reward: 876.59
average score: 809.7835714464074
```

Experimental Results and Discussion of bonus parts (Impact of Twin Q-Networks, Target Policy Smoothing, Delayed Policy Update Mechanism, Action Noise Injection) (bonus) (30%)

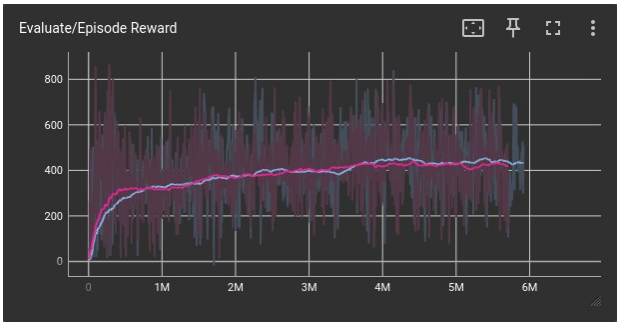
(1) Screenshot of Tensorboard training curve and compare the performance of using twin Q-networks and single Q-networks in TD3, and explain (5%).

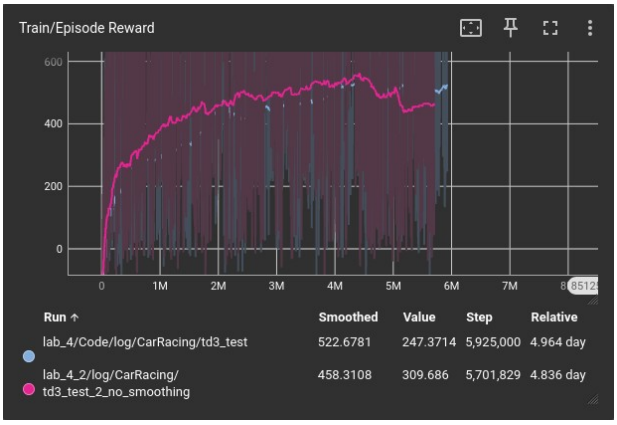




Twin Q-network 在計算目標 Q 值時選擇兩個網路中較小的值（Clipped Double Q-Learning），降低 single Q-network 的 Q 值高估問題。

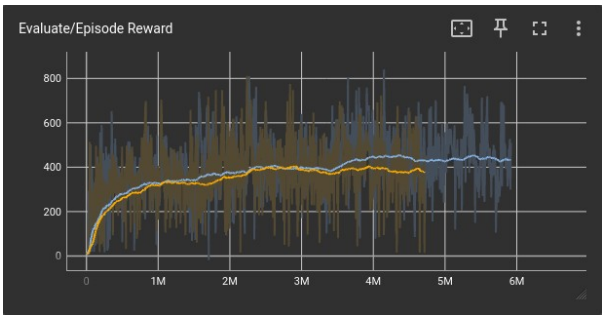
(2) Screenshot of Tensorboard training curve and compare the impact of enabling and disabling target policy smoothing in TD3, and explain (5%).

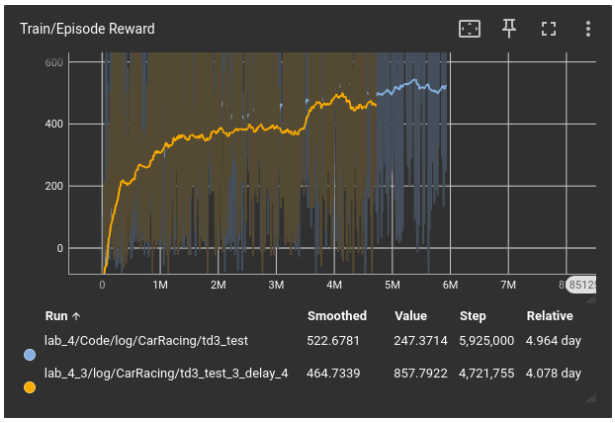




no smoothing 可能導致 Q 值高估、敏感性增加和學習不穩定，在圖片中粉紅色的震盪比原本的幅度大很多，學習曲線相對不穩定。

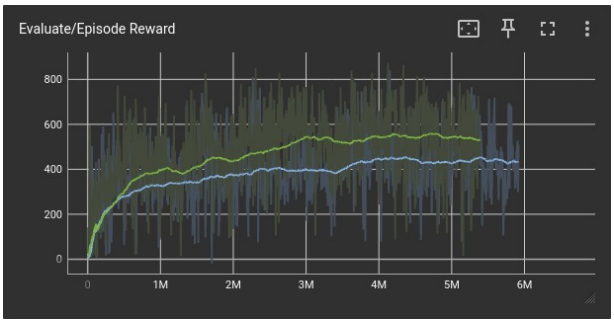
(3) Screenshot of Tensorboard training curve and compare the impact of delayed update steps and compare the results, and explain (5%).

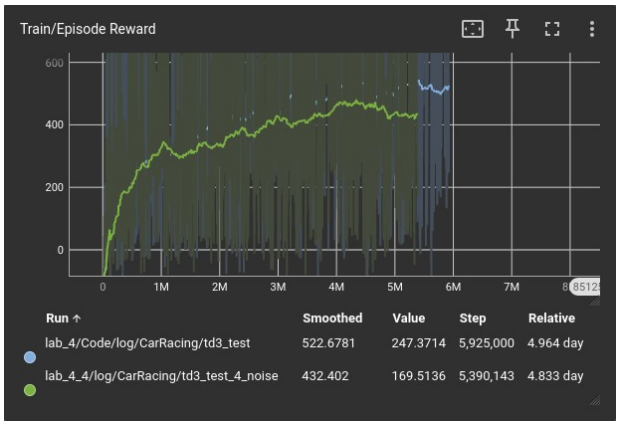




每四次更新可能可以在長期提升穩定性，但可能過於保守，導致早期學習進展稍慢。

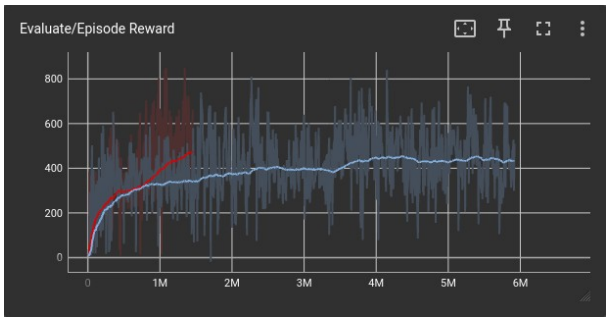
(4) Screenshot of Tensorboard training curve and compare the effects of adding different levels of action noise (exploration noise) in TD3, and explain (5%).

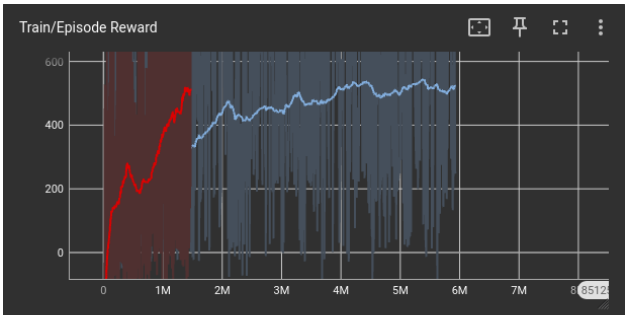




在這個比較中看起來綠色的 OU noise 是比較合適的，可能的原因是因為 OU noise 具有 temporal correlation，具有平滑與連續性。

(5) Screenshot of Tensorboard training curve and compare your reward function with the original one and explain why your reward function works better (10%).






```
# Reward for staying on the road
reward += road_pixel_count * 0.001

# Reward for staying on the grass
reward -= grass_pixel_count * 0.005
```

以上兩個是我對於新的 reward 做的更動，期待的結果是給在路上鼓勵，碰到草地則懲罰，根據實驗也確實學的更快一些。