For RL settings, the hyperparameters are listed in Table I.

TABLE I. Hyperparameters for RL settings

| Parameter for RL | Type/value |
|---|---|
| Optimizer | Adam |
| Activation | ReLU |
| Number of hidden units per layer of policy network | 256/128 |
| Number of hidden units per layer of Q-value network | 256/64 |
| Learning rates of policy and Q-value networks | 5e-5, 1e-4 |
| Minibatch size | 384 |
| Number of epochs | 8 |
| GAE parameter | 0.95 |
| clip epsilon | 0.2 |
| Discount factor | 0.99 |

The training process of RL method based on PPO is shown in Fig. 1.

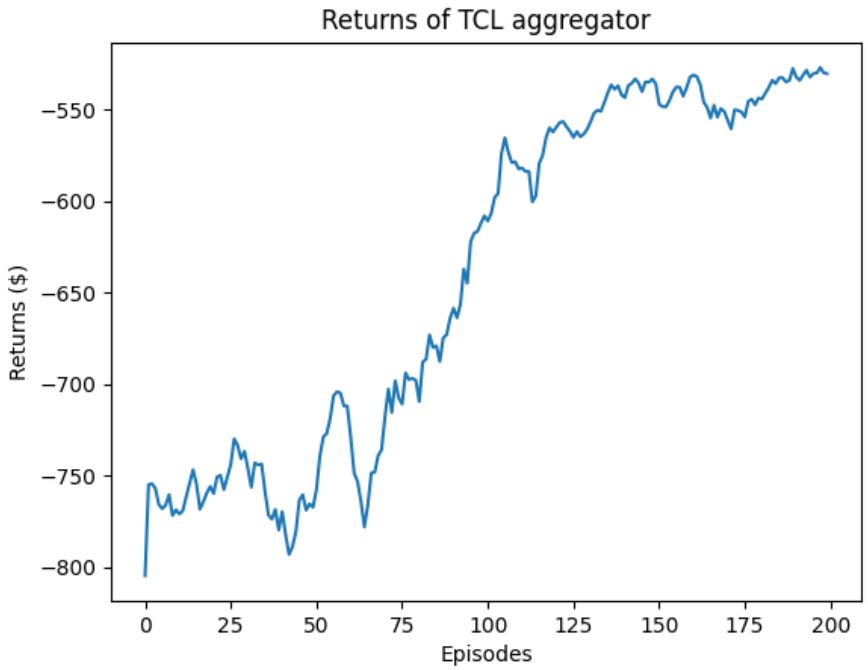


Fig.1 The training process of PPO algorithm