Supplemental Document
Multi-Agent Reinforcement Learning for Autonomous Multi-Satellite Earth Observation: A Realistic Case Study

Author information scrubbed for double-blind reviewing

No Institute Given

1 Multi-Agent Reinforcement Learning Paradigm

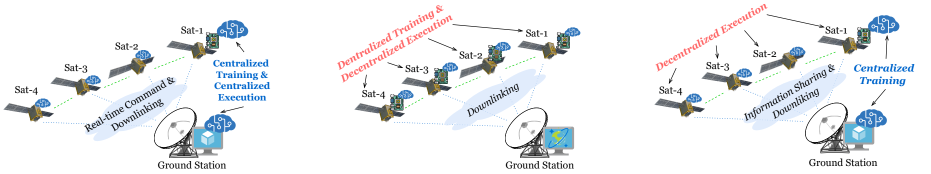
The use of multiple satellites enhances the capabilities of EO missions. These satellites can be viewed as a multi-agent system, particularly when they collaborate to achieve mission objectives. A multi-agent system is commonly modeled as a *Multi-Agent Markov Decision Process (MA-MDP)* or a *Stochastic Game* [1], which is defined by the tuple:

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{r_i\}_{i\in\mathcal{N}}, \mathcal{T}, \gamma \rangle, \tag{1}$$

where $\mathcal{N}$ represents the set of n agents, $\mathcal{S}$ is the state space of the environment, and $\mathcal{A}_i$ is the action space of agent-i, with the joint action space defined as $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$. The state transition probability function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ defines the likelihood of transitioning between states given a joint action. The reward function $r_i : S \times A \rightarrow \mathbb{R}$ determines the individual reward received by each agent-i.

Fully Centralised In this framework, PPO algorithm is used as the single policy network with joint observations as the agent input and the output is the joint actions for each agents. The extension of single-satellite to multi-satellite by using single PPO algorithm is straightforward and less-effort. However, the fully-centralized setting falls easily to sub-optimal point because this method suffers from the non-stationarity issue in cooperative multi-agent settings [2].

Fully Decentralized This MARL framework learns each agent independently and executes its policy based on its local observations and rewards, without any global state or centralized coordination. This is often referred to as Independent MARL (I-MARL) because each agent operates as though it is solving a separate, independent reinforcement learning problem. Each agent optimizes its local policy $\pi_{\theta,i}(a_i, o_i)$ to maximize its expected cumulative reward: $J_i = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(o_{i,t}, a_{i,t})\right]$, where $\mathcal{R}_i(o_{i,t}, a_{i,t})$ is the local reward for agent-i based on its local observation and action. Decentralized training scales naturally with the number of satellites. It is also suitable for environments where satellites


Sat-1
Sat-2
Sat-3
Sat-4
Centralized Training & Centralized Execution
Real-time Command & Downlinking
Ground Station
Dentralized Training & Decentralized Execution
Downlinking
Decentralized Execution
Centralized Training
Information Sharing & Downliking

(a) Fully Centralized (b) Fully Decentralized (c) CTDE

Fig. 1: Three MARL learning frameworks for multi-satellite autonomous EO missions: (a) *Fully Centralized (CTCE)* relies on real-time communication, where all satellite data is transmitted to a central controller—either a ground station or a master satellite with superior computing resources—while other satellites function as slaves. (b) *Fully Decentralized (DTDE)* operates without communication for both training and execution, relying only on local satellite information. On-board AI handles decision-making, with communication limited to EO data downlinking and mission monitoring. (c) *Centralized Training, Decentralized Execution (CTDE)* balances both approaches, keeping training centralized while allowing satellites to execute independently, reducing the need for real-time communication except during training and data downlinking.

have limited communication capabilities or operate in highly dynamic and localized settings, hindering the agent's global information sharing. However, the lack of explicit coordination mechanisms often limit its effectiveness in highly cooperative tasks [3].

2 Parameters for Basilisk Simulator Environment

Several satellite parameters to define the satellite specifications are adjustable from the BSK-RL's environment definition as shown in Table 1.

The orbital parameters in BSK-RL used for our experiments are shown in Table 2.

3 Single Satellite Results

Limited Resources Capacity In single-satellite EO mission, the main resources considered in this study are the battery capacity, data storage capacity, transmission baud rate, and captured image sizes. In our experiments, to demonstrate the resources availability problem the battery and data storage capacity has been defined as: $B = (50, 400)$ Wh, of the Battery and $D = (5, 500)$ GB of the Data Storage. For the transmitter baud rate it is defined as $Bdr = (0.5, 4.3)$ Mbps and captured image size $img = (Small(S), Large(L))$. As shown in Fig 3 left side, generally, the higher amount of resources capacity has less challenge to the PPO learning performance. The limited data storage resources reduces the

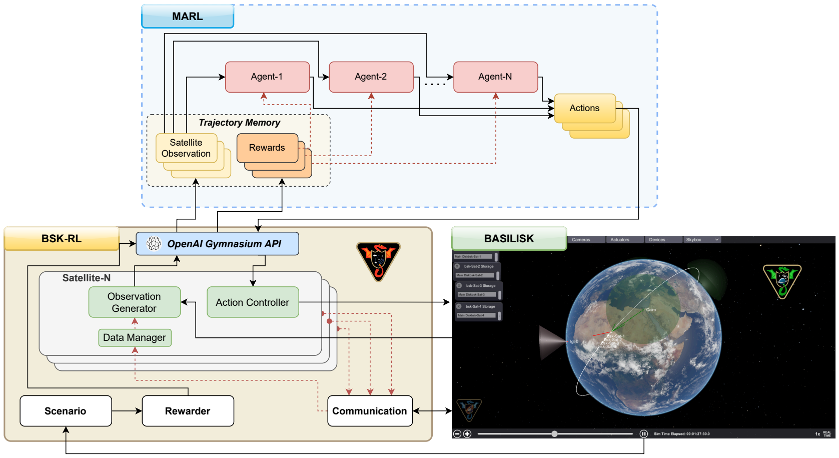


Fig. 2: MARL Framework for Multi-satellite EO Mission Implemented in BSK-RL and the Basilisk Simulator [4] (A realistic satellite system simulator with 3D visualization (Vizard)).

learning performance more than others. A limited battery resource causes multiple significant drops in the learning performance, since it triggers the failure penalty.

Uncertainty and Randomness Challenge Our experiments evaluate the learning performance of PPO under various sources of uncertainty and randomness (see Fig. 3 right side). The presence of randomness and uncertainty introduced fluctuations in PPO's learning process. Several factors of randomness has been set in our experiments: Randomize normally the Attitude (Att.) Disturbance with the scale is 10^{-4} for three attitude angles (yaw, pitch, roll), randomize uniformly the initial Reaction Wheels (RWs) speed in range (-3000,3000) Rpm, the initial battery level in range (40,80)% and the initial data storage level in range (20,80)%. Random initialization of the reaction wheel speed significantly drops the satellite performance to execute accurate maneuvers. Variations in the initial data storage capacity introduced challenges in decision-making and fluctuates the learning performance. Meanwhile, the randomness in the initial battery level has a comparatively smaller impact. The presence of random disturbance in satellite attitude were less impacted since it only affects the capturing performance without any failure penalties. The biggest challenges rise in scenarios with high randomness in reaction wheel speed and data storage, as they significantly increased the difficulty of policy convergence.

Table 1: Satellite Default Parameters

Parameter Name	Default Value	Unit
Data Storage and Transmission:		
Data Storage Capactity	500	GB
Initial Data Storage Level	0	%
Instrument Baud Rate	1000	kbps
Transmitter Baud Rate	4.3	Mbps
Power System:		
Battery Capacity	400	W.h
Initial Battery Level	100	%
Solar Panel Area	1	m^2
Solar Panel Efficiency	20	%
Base Power Draw	-10	W
Instrument Power Draw	-30	W
Thruster Power Draw	-80	W
Satellite Attitude:		
Image Attitude Error	0.1	Degree
Image Rate Error	0.1	Degree
Attitude Control Input Max.	0.4	-
Attitude Rate Control Input Max.	0.1	-
Servo P constant	150	-
Servo Ki constant	5	-
Disturbance vector	[0, 0, 0]	-
Max. Reaction Wheel speed	6000	RPM
Initial Reaction Wheel speed	[0, 0, 0]	RPM

4 Diverse (Heterogeneous) Resource Specification Learning Performance and Coordination

In the diverse resource specification the learning curve are presented in Fig. 4. As the additional results the diverse Battery Capacity satellite action frequencies is shown in Fig. 5 to demonstrates more challenging coordination situation. Although, the policy has coordination behavior, there are some duplications in target captures between the satellites. The battery failure condition is triggered while the battery is empty returning the penalty for the satellite agent. This situation gains more complexity for the satellites to learn from the global reward.

References

1. J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of machine learning research*, vol. 4, no. Nov, pp. 1039–1069, 2003.
2. C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI/IAAI*, vol. 1998, no. 746-752, p. 2, 1998.
3. J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," *arXiv preprint arXiv:1709.04326*, 2017.

Table 2: Orbital Parameters

Parameter	**Single-Sat.**	**Walker-Delta**	**Cluster**
Num. of Satellites	1	4	4
Inclination (deg.)	50	50	50
Offset (deg.)	225	n/a	$225{+}10^{-4}$
Num. of Planes	1	4	1
Altitude (km)	500	500	500

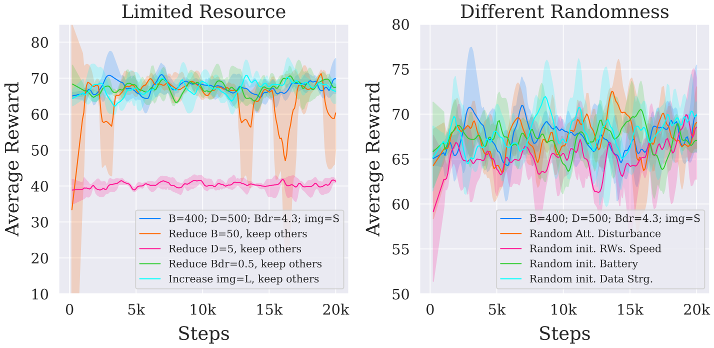


Fig. 3: Single-Satellite Performance Across Various Factors: Battery (B), Data Storage (D), Baudrate (Bdr), and Randomness (see Section 3 for more details).

4. M. A. Stephenson and H. Schaub, "Bsk-rl: Modular, high-fidelity reinforcement learning environments for spacecraft tasking," in *75th International Astronautical Congress, Milan, Italy, IAF*, 2024.

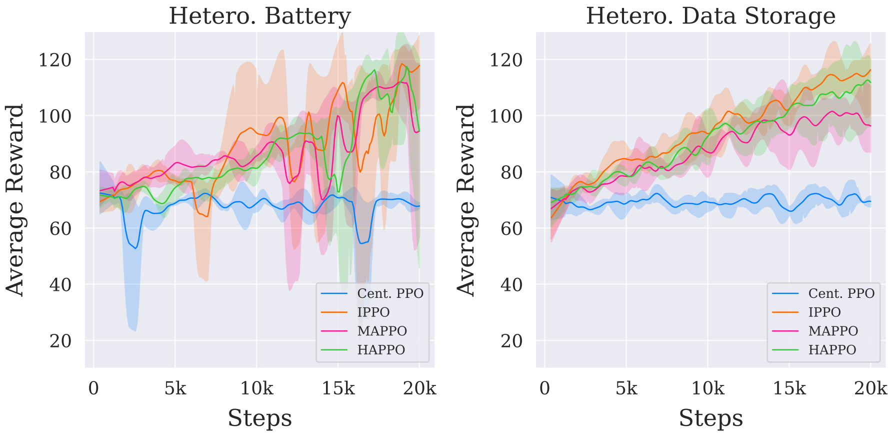


Fig. 4: Multi-Satellite Performance in Diverse (Heterogeneous) Resource Specifications.

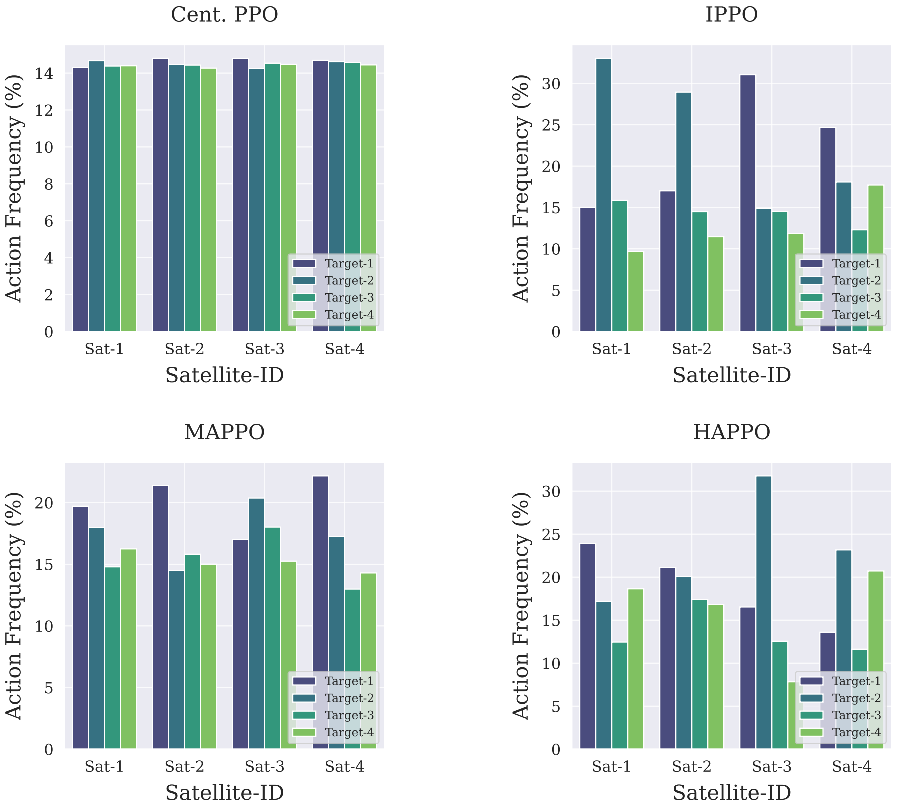


Fig. 5: Target Capturing Action Frequencies Across Different Satellites and Algorithms: Evaluated under varying Battery capacities (B): (Sat-1, Sat-2, Sat-3, Sat-4) =(50, 100, 200, 400) Wh.