


## Article

# The Maintenance of Orbital States in a Floating Partial Space Elevator Using the Reinforcement Learning Method <sup>†</sup>

Weili Xu <sup>1,2</sup>, Xuerong Yang <sup>1,2,\*</sup>  and Gefei Shi <sup>1,2,\*</sup>

<sup>1</sup> School of Aeronautics and Astronautics, Sun Yat-sen University, Guangzhou 510275, China; xuwli6@mail2.sysu.edu.cn

<sup>2</sup> School of Aeronautics and Astronautics, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China

\* Correspondence: yangxr23@mail.sysu.edu.cn (X.Y.); shigf@mail.sysu.edu.cn (G.S.)

<sup>†</sup> This article is an expanded version of a paper entitled Orbital States Keeping of the Floating Partial Space Elevator Using Reinforcement Learning Method. In Proceedings of the 7th International Conference on Tethers in Space, Toronto, ON, Canada, 2–5 June 2024.

**Abstract:** A partial space elevator (PSE) is a multi-body tethered space system in which the main satellite, typically an ultra-large spacecraft or a space station in a higher orbit, is connected to a transport spacecraft in a lower orbit via a tether, maintaining orbital synchronization. One or more climbers can move along the tether driven by electric power, enabling cross-orbital payload transportation between the two spacecraft. The climbers' motion significantly alters the main satellite's orbital states, compromising its safe and stable operation. The dynamic coupling and nonlinearity of the PSE further exacerbate this challenge. This study aims to preliminarily address this issue by proposing a new mission planning strategy. This strategy utilizes reinforcement learning (RL) to select the waiting interval between two transfer missions, thereby maintaining the main satellite's orbital motion in a stable state. Simulation results confirm the feasibility and effectiveness of the proposed mission-based method.

**Keywords:** partial space elevator; maintenance of orbital states; reinforcement learning; mission planning



**Citation:** Xu, W.; Yang, X.; Shi, G. The Maintenance of Orbital States in a Floating Partial Space Elevator Using the Reinforcement Learning Method. *Aerospace* **2024**, *11*, 855. <https://doi.org/10.3390/aerospace11100855>

Academic Editor: Vladimir S. Aslanov

Received: 20 September 2024

Revised: 14 October 2024

Accepted: 15 October 2024

Published: 16 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

A PSE is a multi-body tethered space system (TSS), with its overall structure depicted in Figure 1 [1]. The main satellite, which is typically a space station or ultra-large space structure, is located in a higher orbit [2] and utilizes a tether to anchor the end body, which is typically a transport spacecraft or space experimental platform in a lower orbit. The length of the tether can range from tens to thousands of kilometers, depending on the mission requirements and background. Climbers can move along the tether rapidly using electric power, enabling efficient cross-orbit payload transportation between the two spacecraft. The energy cost can be reduced to less than 60% of traditional rocket transportation means [3], and the unit cost of payload transportation is expected to be reduced to below 5% of traditional means [4]. Additionally, a PSE allows for flexible deployment and deconstruction by releasing or retrieving the tether, following environmental conditions and requirements. Here, it should be noted that in practical applications, tethers several tens of kilometers in length have been used for deep-sea exploration and mooring floats [5,6]. Currently, there are no applications or preparations of tethers that are thousands of kilometers long for ground missions. The weightless environment in which PSEs and space elevators will be located will be an ideal scenario for their future application [7].

In the past decade, studies in the field of tethered spacecraft have mainly focused on the precise modeling, dynamic calculations, and libration suppression of PSEs [8–11]. Works have made initial progress in addressing the challenges related to accurate modeling and computation and have proposed relatively comprehensive libration suppression methods. As research deepens, the way to deal with the orbital floating and oscillation of the main

satellite caused by payload transportation is becoming a key challenge [12,13] that relates to previous work.

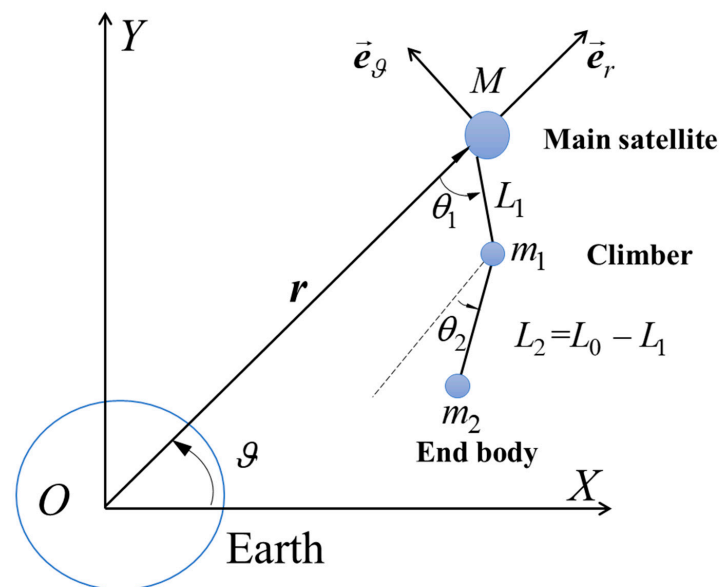


Figure 1. Diagram of a PSE.

The dynamics and modeling of PSEs can include studies on three-body TSSs, which have moving masses. Lorenzini [14], Misra [15], Williams [16], Cohen [17], and Jung [17] et al. built dynamic models of three-body TSSs, and some dynamic characteristics have also been analyzed. In 2010, Woo and Misra [18] introduced the concept of the PSE. Then, the dynamic responses of a PSE to a climber's motion were analyzed by Shi and Zhu et al. [9,19] using a simplified model. These responses were verified by Yu et al. [20] through ground experiments. To provide a more accurate dynamic model of a PSE, Li and Zhu [21,22] developed a nodal position finite element method-based high-fidelity model. This model incorporates most physical factors, such as tether elasticity and flexible motions. They found that the simplified model of a PSE is sufficiently accurate for describing libration and orbital motions. Additionally, Shi and Zhu [12,13] found that the climber's movement along the tether significantly alters the main satellite's orbit, which is noteworthy for maintaining the satellite's orbital stability during payload transportation.

Stability control of PSEs is crucial, primarily focusing on suppressing libration. Consequently, many stabilization techniques for TSSs can be applied to PSEs. Super-twisting sliding mode control laws and MPC methods [23–25] have been investigated to mitigate the libration motion of TSSs, and these control methods are also applicable to PSEs. To achieve a stable configuration in the transfer period of PSEs, the prescribed performance law was considered in a dual-loop control scheme [26]. Optimal control strategies are also effective and desired to suppress the libration from the planning perspective, in general [27,28]. However, the calculation process is complex, and the calculation workload is great. These issues have adverse effects on online control. To deal with these issues, piece-wise optimal schemes [29] and cooperative game methods [1] have been considered. In summary, both the main satellite and the end body need to be controlled, but they are two issues for PSEs. For the control of the main satellite, the target is to maintain its orbital state, especially the orbital radius in a desired realm. For the end body, the target is to maintain its position in a desired realm relating to the main satellite. Since the end body is commonly a transporting spacecraft with high mobility, the control of end bodies is a common issue. However, for the main satellite, which is typically regarded as an ultra-large space structure with a very large mass, its orbital control is a complex issue that has to be considered based on multiple aspects.

In this work, we focus on the maintenance of the main satellite's orbital states. The aforementioned approaches can maintain the orbital states of the main satellite. However, they require additional thrusters and have a limited adjustment range. Our previous research [12] demonstrated that varying the transportation start time, which corresponds to the initial state of the PSE, results in different orbital states after the transportation. This provides a new mission-based planning method for maintaining the orbital states of the main satellite using the exhaustion method. Its drawbacks include (i) its inability to quickly achieve mission planning in orbit with small step sizes and (ii) its omission of optimal and suboptimal solutions with larger step sizes. To solve these issues, in this study, we propose and implement a mission-based RL method by selecting a waiting interval between two transfer missions (one upward and one downward). This new mission-based planning strategy leverages the RL method. The planning policy is trained using the deep Q-network algorithm, where the waiting interval serves as the agent's action. To effectively train the agent, we designed a new reward function that evaluates the main satellite's stability after the climber's transportation. The proposed method can maintain the orbit radius of the main satellite to a large extent without using thrust, thereby saving fuel for the precise orbit control of the main satellite that requires thrust in the future. The numerical simulation results demonstrate the effectiveness of the new mission-based method using RL.

This paper is organized as follows. In Section 2, we model the dynamics of a PSE. Subsequently, the problem is described, and an RL-based method based on the specificity of the problem is proposed in Section 3. The proposed mission-based method is validated by the case study in Section 4. Finally, in Section 5, the conclusions are given in summary, and prospects for future work are provided.

## 2. Mathematical Formulation

As illustrated in Figure 1,  $M$ ,  $m_1$ , and  $m_2$  represent the masses of the main satellite, climber, and end body, respectively. The PSE operates in an ideal central gravitational field with disturbances like solar radiation pressure and atmospheric drag being ignored. The main satellite, which is typically a large-mass space structure, is located in a higher orbit and utilizes a tether to anchor the end body in a lower orbit. The length of the tether can range from tens to thousands of kilometers, depending on the mission requirements and background. The climber can move along the tether rapidly using electric power, enabling efficient cross-orbit payload transportation between the main satellite and the end body.

Here, it should be noted that the out-of-plane motions of the PSE are neglected, as it can be safely decoupled from the in-plane motions due to the weak coupling nature between these two modes of libration. Thus, we focus on the motions in the orbital plane using an in-plane model.

The vectors  $\mathbf{r}$ ,  $\mathbf{r}_1$ , and  $\mathbf{r}_2$  are the projections of the absolute position of the main satellite, the climber, and the end body in the orbital coordinate frame, respectively, such that

$$\begin{aligned} \mathbf{r} &= r\mathbf{e}_r \\ \mathbf{r}_1 &= \mathbf{r} - L_1 \cos \theta_1 \mathbf{e}_r - L_1 \sin \theta_1 \mathbf{e}_\theta \\ \mathbf{r}_2 &= \mathbf{r}_1 - L_2 \cos \theta_2 \mathbf{e}_r - L_2 \sin \theta_2 \mathbf{e}_\theta \end{aligned} \quad (1)$$

where  $r$  denotes the orbital radius of the main satellite,  $\vartheta$  is the true anomaly,  $L_1$  and  $L_2$ , are the tether lengths, and  $\theta_1$  and  $\theta_2$  denote the libration angles of the climber and the end body, which are measured from vector  $\mathbf{r}_0$  to  $L_1$  and  $L_2$ , respectively. The unit vectors  $\mathbf{e}_\theta$  and  $\mathbf{e}_r$  are in the directions of the main satellite's orbital motion and radius, respectively.

Differentiating Equation (1) yields

$$\mathbf{v} = \frac{dr}{dt} \mathbf{e}_r + \omega \times \mathbf{r}, \quad \mathbf{v}_1 = \frac{dr_1}{dt} \mathbf{e}_r + \omega \times \mathbf{r}_1, \quad \mathbf{v}_2 = \frac{dr_2}{dt} \mathbf{e}_r + \omega \times \mathbf{r}_2 \quad (2)$$

where  $\omega = [0 \quad 0 \quad \dot{\vartheta}]^T$  and  $\dot{\vartheta}$  is the angular velocity of the true anomaly. We define the Lagrange equation as  $L = K - U$ , where potential energy ( $U$ ) and kinetic energy ( $K$ ) are

$$\begin{aligned}
 U &= -\mu \left( \frac{M}{|r|} + \frac{m_1}{|r_1|} + \frac{m_2}{|r_2|} \right) \\
 K &= \frac{1}{2} M \mathbf{v}^T \cdot \mathbf{v} + \frac{1}{2} m_1 \mathbf{v}_1^T \cdot \mathbf{v}_1 + \frac{1}{2} m_2 \mathbf{v}_2^T \cdot \mathbf{v}_2
 \end{aligned}
 \tag{3}$$

The generalized coordinates are  $(q_1, q_2, q_3, q_4, q_5, q_6) = (\theta_1, \theta_2, L_1, L_2, r, \dot{\theta})$ , such that the Lagrange equation can be written as

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = Q_i \quad (i = 1, 2, 3, 4, 5, 6)
 \tag{4}$$

The generalized force  $Q_i$  can be derived by

$$Q_i = \sum_{j=0}^2 \Lambda_j^T \frac{\partial \mathbf{r}_k}{\partial q_i} \quad (j = 0, 1, 2)
 \tag{5}$$

where

$$\begin{aligned}
 \Lambda_0 &= -T_1 \cos \theta_1 \mathbf{e}_r - T_1 \sin \theta_1 \mathbf{e}_\theta \\
 \Lambda_1 &= (T_1 \cos \theta_1 - T_2 \cos \theta_2) \mathbf{e}_r + (T_1 \sin \theta_1 - T_2 \sin \theta_2) \mathbf{e}_\theta \\
 \Lambda_2 &= T_2 \cos \theta_2 \mathbf{e}_r + T_2 \sin \theta_2 \mathbf{e}_\theta
 \end{aligned}
 \tag{6}$$

The dynamics of PSE can be derived using the Lagrange method [21] as

$$\ddot{\theta}_1 = -\frac{3\dot{\theta}^2 \sin 2\theta_1}{2} + \frac{T_1 \sin \theta_1}{rM} - \frac{T_2 \sin(\theta_1 - \theta_2)}{L_1 m_1} - \frac{2(\dot{\theta} + \dot{\theta}_1) \dot{L}_1}{L_1} + \frac{2\dot{\theta} \dot{r}}{r}
 \tag{7}$$

$$\ddot{\theta}_2 = -\frac{3\dot{\theta}^2 \sin 2\theta_2}{2} + \frac{T_1 \sin \theta_1}{rM} + \frac{T_1 \sin(\theta_1 - \theta_2)}{(L_0 - L_1) m_1} + \frac{2(\dot{\theta} + \dot{\theta}_2) \dot{L}_1}{L_0 - L_1} + \frac{2\dot{\theta} \dot{r}}{r}
 \tag{8}$$

$$\ddot{L}_1 = \dot{\theta}^2 L_1 - \frac{\mu \cos \theta_1}{r} - \frac{G_1^1 \cos \theta_1 + G_1^2 \sin \theta_1}{m_1} - \frac{T_1}{m_0} - \frac{T_1}{m_1} + \frac{\cos(\theta_1 - \theta_2) T_2}{m_1} + (2\dot{\theta} + \dot{\theta}_1) L_1 \dot{\theta}_1
 \tag{9}$$

$$\begin{aligned}
 \ddot{L}_2 &= \dot{\theta}^2 L_2 + \frac{G_1^1 \cos \theta_2 + G_1^2 \sin \theta_2}{m_1} - \frac{G_2^1 \cos \theta_2 + G_2^2 \sin \theta_2}{m_1} - \frac{T_1}{m_1} - \frac{T_2}{m_2} \\
 &+ (2\dot{\theta} + \dot{\theta}_2) L_2 \dot{\theta}_2 + \frac{T_1 \cos(\theta_1 - \theta_2)}{m_1}
 \end{aligned}
 \tag{10}$$

$$\ddot{r} = r \dot{\theta}^2 - \frac{\mu}{r^2} - \frac{T_1 \cos \theta_1}{M}, \quad \ddot{\theta} = -\frac{2\dot{\theta} \dot{r}}{r} - \frac{T_1 \sin \theta_1}{rM}
 \tag{11}$$

$$\begin{aligned}
 G_1^1 &= -\frac{\mu m_1 (r - L_1 \cos \theta_1)}{|r_1|^3}, & G_1^2 &= \frac{\mu m_1 L_1 \sin \theta_1}{|r_1|^3} \\
 G_2^1 &= -\frac{\mu m_2 (r - L_1 \cos \theta_1 - L_2 \cos \theta_2)}{|r_2|^3}, & G_2^2 &= \frac{\mu m_1 (L_1 \sin \theta_1 + L_2 \sin \theta_2)}{|r_2|^3}
 \end{aligned}
 \tag{12}$$

where  $\mu$  is the Earth's gravitational parameter and  $T_1$  and  $T_2$  are the tether tensions in  $L_1$ .

The tensions in tethers can be calculated by assuming that the total tether length  $L_0$  is constant, such that

$$L_2 = L_0 - L_1, \quad \dot{L}_2 = -\dot{L}_1, \quad \ddot{L}_1 = -\ddot{L}_2
 \tag{13}$$

Substituting Equation (13) into Equations (9) and (10) yields

$$\begin{aligned}
 T_1 &= m_0 \left\{ r^2 G_1^1 [m_2 \sin \theta_2 \sin(\theta_1 - \theta_2) - m_1 \cos \theta_1] \right. \\
 &\quad \left. - r^2 G_1^2 [m_1 \sin \theta_1 + m_2 \cos \theta_2 \sin(\theta_1 - \theta_2)] \right. \\
 &\quad \left. + m_1 \left\{ -\mu (m_1 + m_2) \cos \theta_1 + r^2 \left\{ (m_1 + m_2) \left[ L_1 (\dot{\theta} + \dot{\theta}_1)^2 - \ddot{L}_1 \right] \right. \right. \right. \\
 &\quad \left. \left. - \cos(\theta_1 - \theta_2) \sin \theta_2 G_2^2 - \cos(\theta_1 - \theta_2) \cos \theta_2 G_2^1 \right. \right. \\
 &\quad \left. \left. + m_2 \cos(\theta_1 - \theta_2) \left[ L_2 (\dot{\theta} + \dot{\theta}_2)^2 + \ddot{L}_1 \right] \right\} \right\} / A
 \end{aligned}
 \tag{14}$$

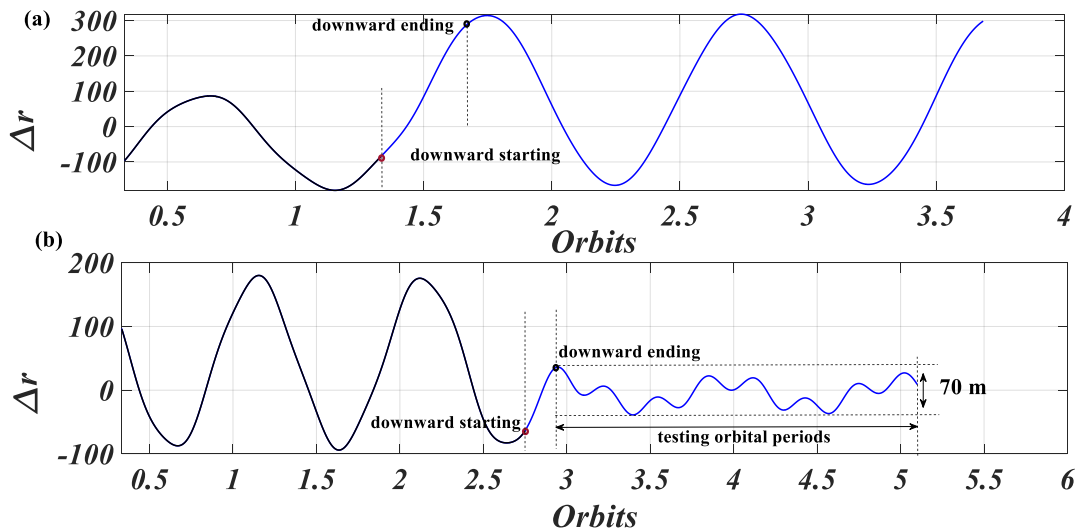
$$\begin{aligned}
 T_2 = & \left\{ -2r^2G_2^1m_1(m_1 + m_0) \cos \theta_2 - 2r^2G_2^2m_1(m_1 + m_0) \sin \theta_2 \right. \\
 & + m_2 \left\{ 2r^2G_1^1[m_0 \sin \theta_1 \sin(\theta_1 - \theta_2) + m_1 \cos \theta_2] \right. \\
 & + \sin \theta_2 \left\{ 2r^2G_1^2(m_1 + m_0 \cos^2 \theta_1) + 2m_1m_0 \left\{ r^2 \left[ L_1(\dot{\theta} + \dot{\theta}_1)^2 - \ddot{L}_1 \right] - \mu \cos \theta_1 \right\} \right\} \\
 & - 2m_0 \cos \theta_1 \cos \theta_2 \left\{ r^2G_1^2 \sin \theta_1 + m_1 \left\{ \mu \cos \theta_1 - r^2 \left[ L_1(\dot{\theta} + \dot{\theta}_1)^2 - \ddot{L}_1 \right] \right\} \right\} \\
 & \left. \left. + 2r^2m_1(m_1 + m_0) \left[ L_2(\dot{\theta} + \dot{\theta}_2)^2 + \ddot{L}_1 \right] \right\} \right\} / (2r^2A)
 \end{aligned}
 \tag{15}$$

where  $A = m_1(m_1 + m_2) + m_0[m_2 \sin^2(\theta_1 - \theta_2) + m_1]$ .

### 3. Maintenance of the Main Satellite’s Orbital States Using the RL Method

#### 3.1. Problem Formation

As shown in Figure 2, the movement of the climber along the tether causes the orbital radius ( $r$ ) to fluctuate rather than remain constant, and the  $|\dot{r}| \neq 0$  after the transfer mission. Furthermore, large values of  $|\dot{r}|$  will lead to significant fluctuations in  $r$ , which is detrimental to the safety of the main satellite.  $\Delta r = r - r_0$  denotes the changing magnitude of  $r$ , where  $r_0$  is the initial orbital radius of the main satellite. One reasonable condition after one transfer mission is that the  $\Delta r$  is small (see the blue lines in Figure 2), and the orbital states after the transfer mission in (b) are better than those in (a). The system parameters in this case are shown in Table 1. In summary, after one transfer mission, the main satellite’s orbital radius fluctuates around its initial value over time with an obvious changing magnitude of  $r$ . This changing magnitude may either decrease or increase after the next transfer mission depending on the starting time of the next transfer mission.



**Figure 2.** The expected goal to be achieved after the cargo transportation. (a) Next transportation starts without waiting (b) Next transportation starting time obtained by exhaustive method.

Then, the problem is how to select such a mission starting time so as to minimize the changing magnitude of  $r$  after one transfer mission as much as possible. Such selection is difficult. This is because the required input in the mission plan is the mission starting time, a time point. Its dynamic impact is caused by the movement of the climber within a transfer period after that mission starting time. It is difficult to describe the relationship between the mission starting time and the changing magnitude of  $r$  after the transfer period in dynamics. As a result, existing optimization theories are not suitable for optimizing the selection of the mission starting time.

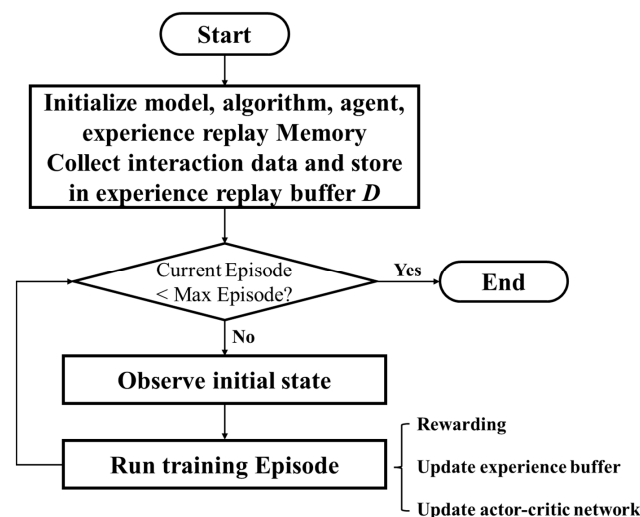
**Table 1.** System parameters and initial states of the PSE.

Parameters	Values
Orbital radius of the main satellite, $r$ (m)	$7.1 \times 10^6$
Initial True anomaly angular, $\vartheta(0)$ (rad)	0
Mass of the main satellite (net weight), $M$ (kg)	50,000
Mass of climber (net weight), $m_1$ (kg)	100
Mass of the payload (kg)	400
Mass of end body, $m_2$ (kg)	1000
Fixed total tether length, $L_0$ (m)	$2 \times 10^4$
Initial libration angle and angular velocity, $(\theta_1(0), \dot{\theta}_1(0))$	(0, 0)
Initial libration angle and angular velocity, $(\theta_2(0), \dot{\theta}_2(0))$	(0, 0)
Initial distance between the climber and the main satellite, $L_1$ (m)	19,500
Reward function parameter, $\lambda$	100
Adaptable average orbital radius changing magnitude $\Delta\bar{r}$ (m)	100

Considering the situations above, one desired strategy is the deep RL method. The RL method for the maintenance of the orbital states of a PSE involves (i) utilizing the powerful fitting capabilities of deep neural networks to establish the mapping relationship between the departure time and the main satellite's orbital state after the transfer mission and (ii) utilizing a trained agent to make decisions. In summary, deep neural networks can be used to address the problem (i), but they must also make decisions on mission execution to provide an action, which is why RL methods are needed. The powerful representation ability of deep neural networks makes them a powerful tool for state representation and function approximation in RL.

### 3.2. RL-Based Mission Planning Method

This work employs the RL method to address the complex dynamics of the PSE by training an agent (controller) using the Deep Q-Network (DQN) learning algorithm. Based on the problem formulation, mission planning for the PSE in this study involves selecting the optimal mission starting time. This can be facilitated by determining the length of the waiting interval  $h$  from the end time point of the previous transfer mission. Then, the action can be defined in a discrete dataset. Once the action is completed, the episode ends. As a result, the length of each episode is equal to  $h$ . The initial states of the agent for each episode are the same. Then, the DQN learning algorithm framework can be summarized as follows (see Figure 3). First, the model, training algorithm, agent (action), and experience replay memory are initialized and set. Interaction data and storage data are collected, then the training begins for the preset episode numbers. The detailed process is as follows:

**Figure 3.** Flowchart of the RL method using the DQN algorithm.

(1) Learning environment: The DQN learning algorithm is designed to solve decision problems based on Markov decision processes (MDPs). The learning agent interacts with an environment, such that:

The state space is defined as the generalized coordinates of the PSE, including  $\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2, L_1, \dot{L}_1, r, \dot{r}, \vartheta, \dot{\vartheta}$ .

The action is defined as the waiting interval  $h$ .

The reward function determines the immediate feedback that guides the agent's learning. The agent aims to maximize the cumulative reward over time, and the specific design of the reward function in this work is defined as

$$Reward = \Delta\bar{r} - (\max|r|_n - \min|r|_n) - \lambda \max|\dot{r}|_n \quad (16)$$

where  $\Delta\bar{r}$  is an average orbital radius changing magnitude which can be set by the user, while  $\lambda$  is a constant parameter.  $\lambda > 0$  is a constant parameter that is used to penalize excessive orbital radius change rates.  $|r|_n$  and  $|\dot{r}|_n$  denote the orbital radius of the main satellite in  $n$  test orbital periods. Here, the test orbital period is defined as the orbital period after the transfer mission (see Figure 2b). In such orbital periods, the climber's transportation is completed, and the main satellite's radius keeps fluctuating.  $\max|r|_n - \min|r|_n$  denotes the amplitude of change in the main satellite's orbital radius over  $n$  test orbital periods after the transfer mission. Here, it should be noted that in this work, each episode comprises only one action.

(2) Construct the Q-network: The network uses generalized coordinates as the input and generates Q-value estimates for each possible action.

(3) Initialize network parameters: The biases and weights of the deep Q-network are initialized.

(4) Select an action: Based on the estimations and current state provided by the deep Q-network, an action is chosen using an  $\epsilon$ -greedy strategy to balance exploration and exploitation.

(5) Execute the action: The chosen action is performed, and the environment will return a reward and the next state.

(6) Record experiences: Current states ( $\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2, L_1, \dot{L}_1, r, \dot{r}, \vartheta, \dot{\vartheta}$ ), actions ( $\Delta t$ ), rewards, and next states are stored as experience tuples.

(7) Experience replay: A batch of experience tuples are randomly sampled from the experience storage to train the deep Q-network.

(8) Update Q-values: The sampled experiences are utilized to adjust the parameters of the deep Q network, aiming to reduce the disparity between the predicted Q-values and the target Q-values.

(9) Compute target Q-values: Target Q-values for updating the deep Q-network are determined. These target Q-values include the highest Q-value for the subsequent state combined with the current reward.

(10) Train the network: Backpropagation and gradient descent are applied to reduce the discrepancy between the predicted Q-values and the target Q-values, thereby updating the parameters of the deep Q network.

(11) Steps (4) to (10) are repeated until a stopping condition or convergence is reached.

By iteratively updating the parameters of the deep Q-network, the DQN learning algorithm gradually improves the estimation of Q-values and discovers the optimal policy for RL missions.

#### 4. Numerical Simulation and Discussion

The proposed mission-based method is validated by the case study with the following system parameters and the initial states of the PSE in Table 1.

The payload is assumed to be 400 kg and the climber's net weight is 100 kg, such that, in the upward transportation,  $M = 50,000$  kg and  $m_1 = 500$  kg. In the waiting interval,  $M = 50,500$  kg, and in the downward transfer period,  $m_1 = 100$  kg and  $M = 50,400$  kg. The

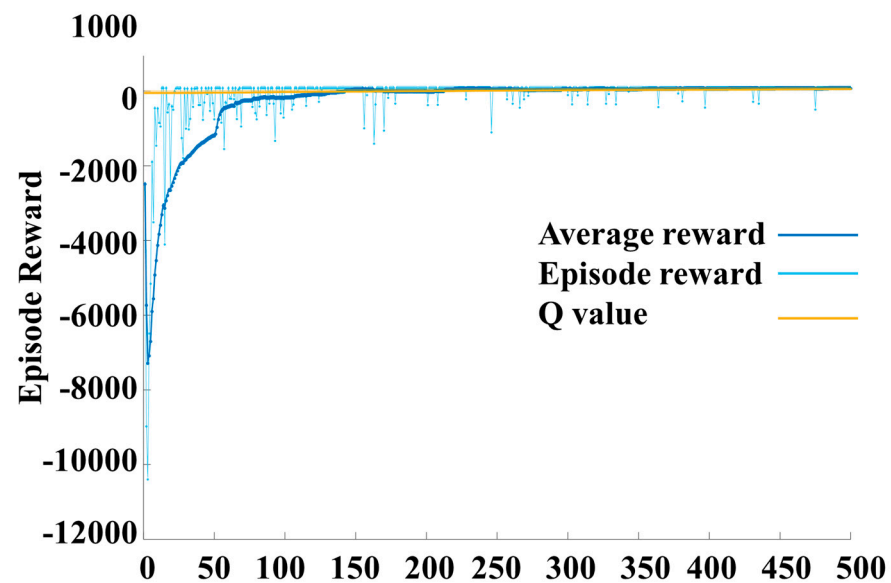
speed function of the climber is designed as a sine form,  $\dot{L}_1 = 4.9\pi \sin(\pi t/2000)$ . The mission plan aims to find an optimal waiting interval  $h$  to minimize  $\Delta r$  after the downward transfer mission. The policy is trained for 500 episodes. The hyperparameters of the DQN are shown in Table 2.

**Table 2.** DQN hyperparameters.

Parameters	Values
Learning rate	0.01
Gradient threshold	inf
Discount factor	0.99
Batch size	100
Experience buffer length	1000

In this work, the  $\epsilon$ -greedy exploration strategy is used with an  $\epsilon$  decay of 0.005, and a minimum  $\epsilon$  set at 0.01, where  $\epsilon$  is the exploration factor. Long short-term memory is adopted to train the recurrent neural network with two fully connected layers employed with 40 hidden nodes, respectively. In the enhancement of the learning process, rectified linear units are incorporated between each layer. Adam is used as the optimizer. In the training process, the episode step is 1/36 of the orbit, which equals the step of  $h$  in the discrete dataset, and the dynamic time step in the RK-4 is 0.001 s. Numerical simulations are performed using MATLAB, specifically the RL App. Here, it should be noted that by using the Reinforcement Learning Designer App in MATLAB, the simulation can be replicated using the given parameters in Tables 1 and 2.

The results of the numerical simulations are presented in Figures 4–7. The variation in the episode reward with the training process throughout the entire learning process is shown in Figure 4. The first 150 episodes mainly represent the stages of algorithm random exploration and the accumulation of learning samples, corresponding to low reward values, and most of the tasks in the generated plan cannot be completed. As the learning process progresses, the RL method (using the DQN algorithm) gradually summarizes the characteristics of the planning model and obtains a higher return on the reward value. After 200 episodes of learning, the algorithm converges, and the reward value remains stable, at approximately 128.79.



**Figure 4.** Training episodes' mean reward.



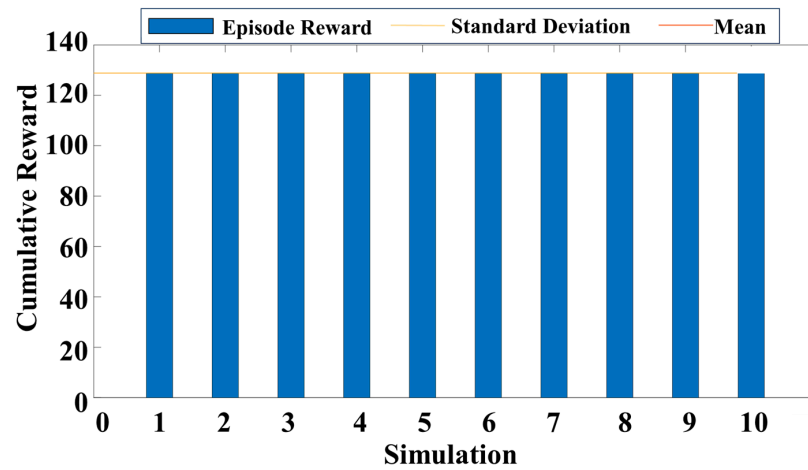


Figure 5. Cumulative reward.

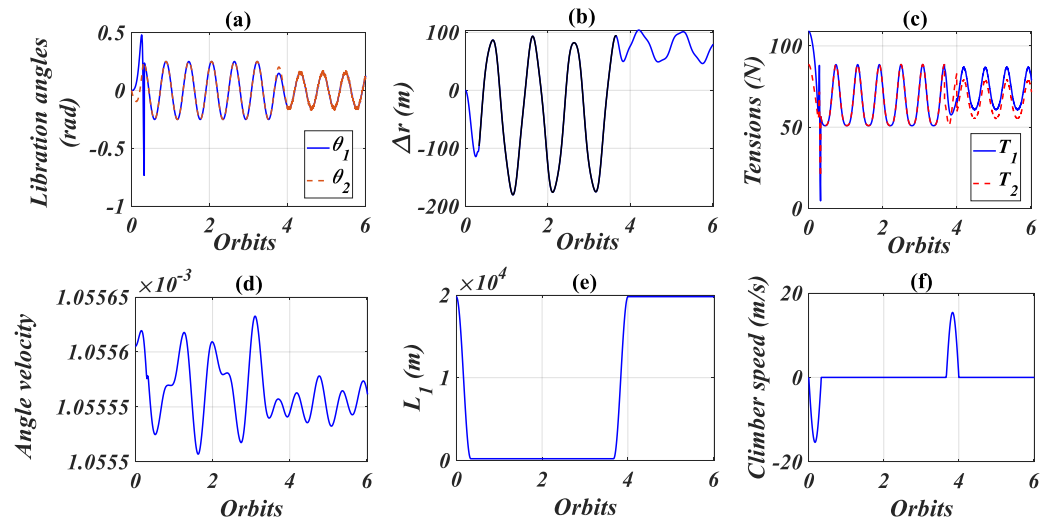
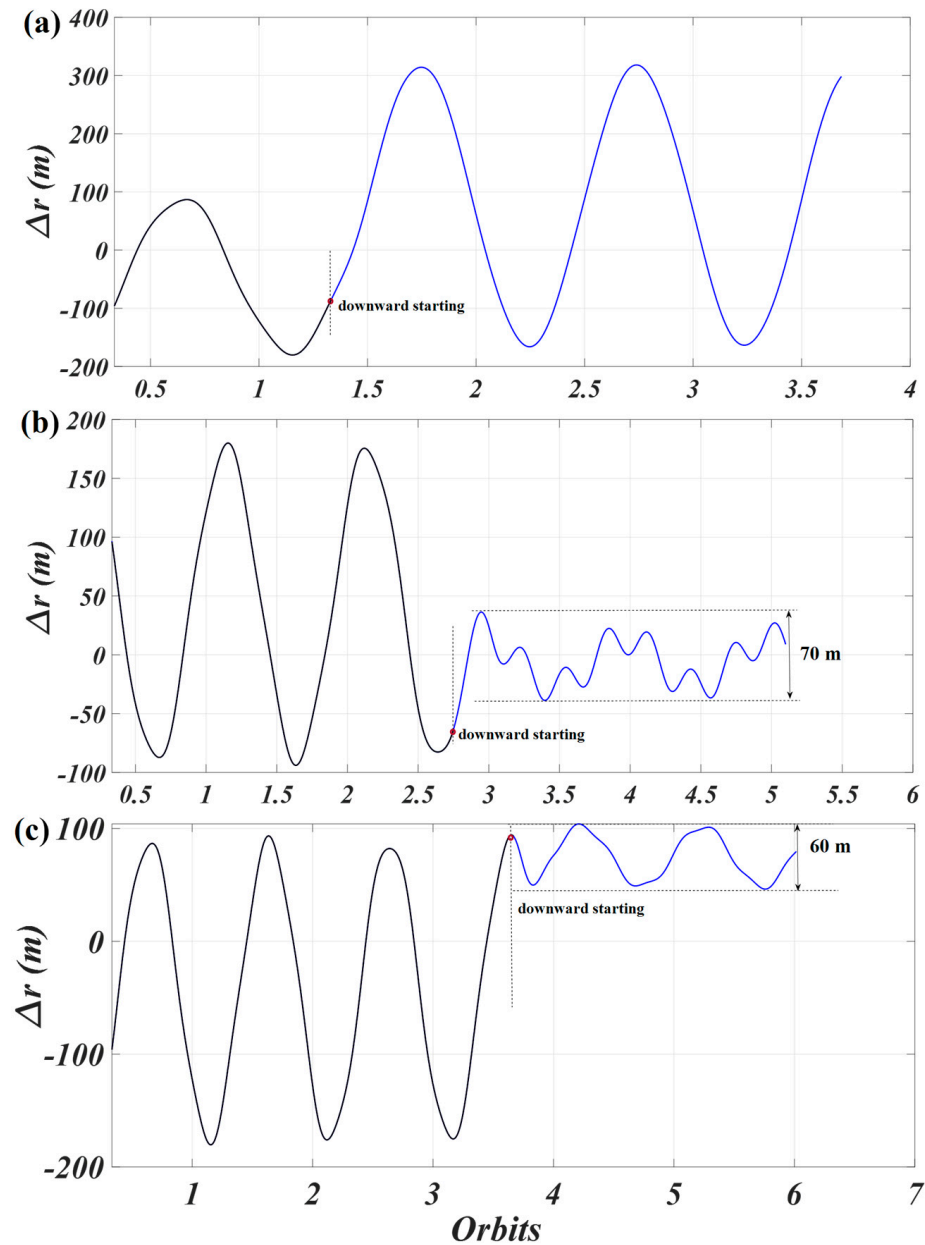


Figure 6. Planning results: (a) libration angles; (b) orbital radius' changing; (c) tensions; (d) main satellite's orbital angular velocity; (e) tether length of  $L_1$ ; (f) climber speed.

The trained agent offers the policy to decide how long the PSE should wait for the next transfer mission. Figures 5 and 6 show the effects of the obtained policy. The waiting interval after an upward transfer mission begins at 2000 s. Here, it should be noted that the randomness of the obtained policy's decisions always exists due to the complexity of the corresponding problem in terms of dynamics and the limited number of training episodes. Thus, a relatively small number of simulations are still needed to demonstrate the results. Thus, the agent gives 10 waiting intervals to achieve the highest reward (see Figure 5). All reward values are near 130 with the waiting interval  $h = 3.67$  orbits = 19,674.7 s.

Figure 6 shows the state of the PSE following the planned mission. The libration angles fluctuate in the mission period by a magnitude of 0.6 rad in the beginning, and then both  $\theta_1$  and  $\theta_2$  change by magnitudes less than 0.3 rad due to the lack of libration suppression (see Figure 6a). The tensions of  $T_1$  and  $T_2$  are greater than zero in the simulation period, which means that the tethers are straight (see Figure 6c). This matches the assumptions in Section 2. The angular velocity in Figure 6d shows that the main satellite runs around the Earth with a nearly constant angular velocity, even though the orbital radius changes in the scale of 100 m. Figure 6e,f show the tether length change for  $L_1$  and the climber speed  $\dot{L}_1 = 4.9\pi \sin(\pi t/2000)$ . This function ensures that the climber moves along the tether following the prescribed distance and period with the initial and final speeds being zero, which matches the engineering condition well. The upward transfer mission leads

to  $\max|r| - \min|r| > 260m$ , as seen in the curve of  $\Delta r$  in the period of 0.4–3.6 orbits in Figure 6b. Then, the downward transfer of the climber following the planned waiting interval reduces the amplitude change of the orbital radius to  $60m$ . This means that the proposed orbital state maintenance method is effective.



**Figure 7.** Comparison of (a) without mission planning, (b) exhaustion, and (c) the RL method.

Figure 7 shows that the waiting interval generated by the proposed RL method reduces the magnitude of orbital radius change after the mission by over 88% compared to the case in which the waiting interval is one orbital period, comparing the blue lines in (a) and (c). Figure 7b shows the result with the waiting interval obtained by exhaustion with the step of 5 min, which is small enough for the exhaustion of a PSE. Although the main satellite's orbital state is kept near zero, the amplitude of the orbital radius is 70 m, which is greater than that in the case in which the waiting interval is generated by the RL method. This is because the searching step of the exhaustion is not small enough to achieve the optimal waiting interval.

## 5. Conclusions

This work presents an RL method to address the maintenance of the orbital states of the main satellite of a PSE based on the requirements and constraints of transportation missions. This new method focuses on planning the waiting interval between two transfer missions from a mission planning perspective in order to minimize the amplitude changes of the orbital radius of the main satellite after completing a transfer mission. The deep Q-network algorithm is employed to train the mission planning policy using a novel reward function. The simulation results demonstrate the following findings:

(i) The proposed mission-based RL method effectively maintains the orbital states of the main satellite.

(ii) With the transportation mission and the proposed actions, even a small number of training episodes can produce effective agents.

(iii) Compared to the exhaustion method, the RL method enables quick decision-making with agents trained offline.

Furthermore, based on the simulation results and conclusions in this work, it can be found that it is necessary to conduct targeted and in-depth research on the dynamic evolution mechanisms of PSEs. Since the overlapping time scales of transfer missions and orbits make the system dynamics mechanisms of the climber transportation mission process unclear, it is difficult to establish effective mapping between “mission orbits” through deep neural networks. This leads to the task sequence planning method based on RL being poorly generalized.

**Author Contributions:** Conceptualization, G.S.; methodology, G.S., W.X. and X.Y.; software, G.S. and W.X.; validation, G.S., W.X. and X.Y.; formal analysis, W.X.; investigation, G.S., W.X. and X.Y.; resources, G.S. and X.Y.; data curation, G.S.; writing—original draft preparation, G.S. and W.X.; writing—review and editing, G.S. and W.X.; visualization, X.Y.; supervision, X.Y.; project administration, G.S.; funding acquisition, G.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is funded by the National Natural Science Foundation of China (Grant No. 12102487, 62388101) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515012339).

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Shi, G.; Zhu, Z.H. Cooperative game-based multi-objective optimization of cargo transportation with floating partial space elevator. *Acta Astronaut.* **2023**, *205*, 110–118. [\[CrossRef\]](#)
2. Yang, G.; Shen, H.; Li, Q.; Wu, S.; Jiang, J. Orbit-attitude-structure-thermal coupled modelling method for large space structures in unified meshes. *Appl. Math. Model.* **2024**, *135*, 26–50. [\[CrossRef\]](#)
3. Woo, P.; Misra, A.K. Energy considerations in the partial space elevator. *Acta Astronaut.* **2014**, *99*, 78–84. [\[CrossRef\]](#)
4. Dixit, U.S.; Dwivedy, S.K.; Forward, T.W. *Mechanical Sciences: The Way Forward*; Springer: Heidelberg, Germany, 2020.
5. Jiang, Y.; Lv, M.; Li, J. Station-keeping control design of double balloon system based on horizontal region constraints. *Aerosp. Sci. Technol.* **2020**, *100*, 105792. [\[CrossRef\]](#)
6. Meng, J.; Zhang, L.; Li, J.; Lv, M. Dynamic modeling and simulation of tethered stratospheric satellite with thermal effects. *Appl. Therm. Eng.* **2017**, *110*, 181–189. [\[CrossRef\]](#)
7. Nixon, A.; Knapman, J.; Wright, D.H. Space elevator tether materials: An overview of the current candidates. *Acta Astronaut.* **2023**, *210*, 483–487. [\[CrossRef\]](#)
8. Jung, W.; Mazzoleni, A.P.; Chung, J. Nonlinear dynamic analysis of a three-body tethered satellite system with deployment/retrieval. *Nonlinear Dyn.* **2015**, *82*, 1127–1144. [\[CrossRef\]](#)
9. Shi, G.; Zhu, Z.; Zhu, Z.H. Libration suppression of tethered space system with a moving climber in circular orbit. *Nonlinear Dyn.* **2017**, *91*, 923–937. [\[CrossRef\]](#)
10. Kojima, H.; Fukatsu, K.; Trivailo, P.M. Mission-function control of tethered satellite/climber system. *Acta Astronaut.* **2015**, *106*, 24–32. [\[CrossRef\]](#)

11. Li, G.; Zhu, Z.H.; Shi, G. A novel looped space tether transportation system with multiple climbers for high efficiency. *Acta Astronaut.* **2021**, *179*, 253–265. [[CrossRef](#)]
12. Shi, G. A mission-based orbit keeping method of the partial space elevator. In Proceedings of the ASCEND 2021, Las Vegas, NV, USA, 15–17 November 2021; AIAA: Reston, VA, USA, 2021.
13. Shi, G.; Zhu, Z.H. Libration-free cargo transfer of floating space elevator. *Nonlinear Dyn.* **2022**, *110*, 2263–2281. [[CrossRef](#)]
14. Lorenzini, E.C.; Cosmo, M.; Vetrella, S.; Moccia, A. Dynamics and control of the tether elevator/crawler system. *J. Guid. Control Dyn.* **1989**, *12*, 404–411. [[CrossRef](#)]
15. Misra, A.K.; Amier, Z.; Modi, V.J. Attitude dynamics of three-body tethered systems. *Acta Astronaut.* **1988**, *17*, 1059–1068. [[CrossRef](#)]
16. Williams, P. Dynamic multibody modeling for tethered space elevators. *Acta Astronaut.* **2009**, *65*, 399–422. [[CrossRef](#)]
17. Cohen, S.S.; Misra, A.K. The effect of climber transit on the space elevator dynamics. *Acta Astronaut.* **2009**, *64*, 538–553. [[CrossRef](#)]
18. Woo, P.; Misra, A.K. Dynamics of a partial space elevator with multiple climbers. *Acta Astronaut.* **2010**, *67*, 753–763. [[CrossRef](#)]
19. Shi, G.; Zhu, Z.; Zhu, Z.H. Dynamics and control of three-body tethered system in large elliptic orbits. *Acta Astronaut.* **2018**, *144*, 397–404. [[CrossRef](#)]
20. Yu, B.S.; Ji, K.; Wei, Z.T.; Jin, D.P. In-plane global dynamics and ground experiment of a linear tethered formation with three satellites. *Nonlinear Dyn.* **2022**, *108*, 3247–3278. [[CrossRef](#)]
21. Shi, G.; Li, G.; Zhu, Z.; Zhu, Z.H. A virtual experiment for partial space elevator using a novel high-fidelity FE model. *Nonlinear Dyn.* **2018**, *95*, 2717–2727. [[CrossRef](#)]
22. Li, G.; Shi, G.; Zhu, Z.H. Three-Dimensional High-Fidelity Dynamic Modeling of Tether Transportation System with Multiple Climbers. *J. Guid. Control Dyn.* **2019**, *42*, 1797–1811. [[CrossRef](#)]
23. Li, X.; Sun, G.; Xue, C. Fractional-order deployment control of space tethered satellite via adaptive super-twisting sliding mode. *Aerosp. Sci. Technol.* **2022**, *121*, 107390. [[CrossRef](#)]
24. Zhang, F.; Huang, P. Releasing Dynamics and Stability Control of Maneuverable Tethered Space Net. *IEEE/ASME Trans. Mechatron.* **2017**, *22*, 983–993. [[CrossRef](#)]
25. Wen, S.; Zhang, F.; Shen, G.; Huang, P. Smooth and Stable Deployment Control of Tether Satellite System using Nonlinear Model Predictive Control With Actuator Constraints. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, 1–10. [[CrossRef](#)]
26. Shi, G.; Zhu, Z.H. Prescribed performance based dual-loop control strategy for configuration keeping of partial space elevator in cargo transportation. *Acta Astronaut.* **2021**, *189*, 241–249. [[CrossRef](#)]
27. Williams, P.; Ockels, W. Climber motion optimization for the tethered space elevator. *Acta Astronaut.* **2010**, *66*, 1458–1467. [[CrossRef](#)]
28. Wen, H.; Zhu, Z.H.; Jin, D.; Hu, H. Tension control of space tether via online quasi-linearization iterations. *Adv. Space Res.* **2016**, *57*, 754–763. [[CrossRef](#)]
29. Shi, G.; Zhu, Z.; Zhu, Z.H. Parallel Optimization of Trajectory Planning and Tracking for Three-body Tethered Space system. *IEEE/ASME Trans. Mechatron.* **2019**, *24*, 240–247. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.