applied sciences

MDPI

Article

Reinforcement-Learning-Based Vibration Control for a Vehicle Semi-Active Suspension System via the PPO Approach

Shi-Yuan Han * and Tong Liang

Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China; liangtong0606@163.com

* Correspondence: ise_hansy@ujn.edu.cn; Tel.: +86-531-8276-6503

Abstract: The vehicle semi-active suspension system plays an important role in improving the driving safety and ride comfort by adjusting the coefficients of the damping and spring. The main contribution of this paper is the proposal of a PPO-based vibration control strategy for a vehicle semi-active suspension system, in which the designed reward function realizes the dynamic adjustment according to the road condition changes. More specifically, for the different suspension performances caused by different road conditions, the three performances of the suspension system, body acceleration, suspension deflection, and dynamic tire load, were taken as the state space of the PPO algorithm, and the reward value was set according to the numerical results of the passive suspension, so that the corresponding damping force was selected as the action space, and the weight matrix of the reward function was dynamically adjusted according to different road conditions, so that the agent could have a better improvement effect at different speeds and road conditions. In this paper, a quarter–car semi-active suspension model was analyzed and simulated, and numerical simulations were performed using stochastic road excitation for different classes of roads, vehicle models, and continuously changing road conditions. The simulation results showed that the body acceleration was reduced by 46.93% under the continuously changing road, which proved that the control strategy could effectively improve the performance of semi-active suspension by combining the dynamic changes of the road with the reward function.

Keywords: proximal policy optimization; vehicle semi-active suspension; road change; reward function

Citation: Han, S.-Y.; Liang, T. Reinforcement-Learning-Based Vibration Control for a Vehicle Semi-Active Suspension System via the PPO Approach. *Appl. Sci.* **2022**, *12*, 3078. https://doi.org/10.3390/app12063078

Academic Editors: Ignazio Dimino and Antonio Concilio

Received: 17 February 2022
Accepted: 15 March 2022
Published: 17 March 2022

1. Introduction

In the Nineteenth Century, the quality of people's travel was greatly improved with the birth of the first automobile. The components of semi-active suspension systems, which mainly include shock absorbers, elastic elements, and dampers, play a key role in improving the smoothness, handling stability, and passenger comfort of a car.

An onboard sensor is used to extract the road information for the semi-active suspension system, through the adjustment of the damping coefficient to improve the smoothness and stability of the vehicle. At present, the control strategy of semi-active suspension has been extensively studied. A controller was proposed by [1] to characterize the inverse dynamics of magnetorheological dampers using a polynomial model. Rao et al. [2] proposed an improved sky-hook control (SHC) strategy to optimize the parameters by equating the control force of the feedback system with that obtained by a linear quadratic regulator (LQR). The paper [3] proved the conditions for the feedback controller in the ideal state using linear matrix inequalities (LMIs) and Liapunov functions that depend on the time lag distance. Based on variational optimal control theory, an open-loop variational feedback controller for semi-active control of a suspension system was proposed in [4]. A fuzzy controller and an adaptive fuzzy controller were developed in [5] to achieve road holding and passenger comfort on most road profiles. The article [6] developed an adaptive fuzzy sliding mode control method based on a bionic reference model, which takes into account the variation of the vehicle mass and is able to provide both finite-time convergence and

energy-efficient performance. The article [7] developed a new dynamic model named the QSHM model based on quasi-static (QS) models and hysteretic multiplier (HM) factors. On the basis of the implementation of a new method of experiment-based modeling of magnetorheological hydrodynamics, several hysteresis multiplier factors were introduced. The article [8] developed a self-powered magnetic rheological (MR) damper to effectively control the vibration of a front-loaded washing machine. The damper consisted of a shear MR damping part and a generator with a permanent magnet and induction coil. The paper [9] proposed a new adaptive control method to deal with dead zones and time delay problems in actuators of vibration control systems. In addition, fuzzy neural networks were used to approximate unmodeled dynamics, and sliding mode controllers were designed to enhance the robustness of the system to uncertainty and robustness. Based on the Bolza–Meyer criterion, a new optimal control law related to sliding mode control was developed in [10]. The controller has gain adjustability, where the gain value can be greater than one. Most of the aforementioned methods are based on establishing an accurate mathematical model and proving the effectiveness of the proposed control strategy through a rigorous theory, which is computationally intensive and more difficult to implement in practice.

Since its introduction in 1954, reinforcement learning has been developed over the years with reformative advances in both theory and applications. In terms of theory, the deep Q-network (DQN) algorithm was proposed in [11], which solved the problem of the difficult convergence of training under off-policy by end-to-end training mode. The deep deterministic policy gradient (DDPG) algorithm was proposed in [12], which is based on the DQN, using a replay buffer and target network for updating, while using a neural network for function approximation. The paper [13] proposed the trust region policy optimization (TRPO) algorithm, which was improved for policy optimization and solves the problem of selecting the update step size during the policy update, which enables the policy update of the agent to be monotonically enhanced, but it also has the problem of high computational complexity. Compared with control strategies based on mathematical models, reinforcement learning can generate a large amount of data by interacting with the environment and find the best strategy from the sampled data, so it has been successfully applied in several application scenarios. In 2017, Alpha Go defeated Ke Jie by using reinforcement learning algorithms, which is the world's top-ranked Go player, and since then, reinforcement learning has been known to the public [14] from the Wuzhen Go Summit. The paper [15] proposed a deep reinforcement learning framework in the MOBA game Honor of Kings, enabling the AI agent Tencent Solo to beat top human professional players in 1v1 games. In recent years, the PPO algorithm [16] based on the actor–critic algorithm uses an importance sample to transform the on-policy into the off-policy, which is conducive to fully reusing the sampled data, avoiding data waste and speeding up the learning rate, and the CLIP function is used to control the step size of the gradient update of the policy, so that the control strategy is monotonically improved without violent shaking. The PPO algorithm has been applied in several application scenarios. In paper [17], the PPO algorithm was applied to image caption generation; a single model was trained by fine-tuning the pre-trained X-Transformer, and good experimental results were obtained. The paper [18] utilized the proximal policy optimization (PPO) algorithm to construct a second-order intact agent for navigating through an obstacle field such that the angle-based formation was allowed to shrink while maintaining the shape, and the geometric centroid of the formation was constantly oriented toward the target. The PPO algorithm is suitable as a control strategy for semi-active suspensions because of its ability to continuously interact with the environment to achieve policy improvement and optimize monotonic enhancement through importance sampling, as well as due to its computational simplicity.

At present, the application of reinforcement learning in vehicle suspension control has attracted much attention and achieved a series of results in order to reduce the dependence on accurate mathematical models for the design of control strategies for suspension systems and to overcome the influence of its own uncertain parameters and the external environment on the control performance. The article [19] explored the application of batch

reinforcement learning (BRL) to the problem of designing semi-active suspensions for optimal comfort. In the article [20], the performance of a nonlinear quarter-car active suspension system was investigated using a stochastic real-valued reinforcement learning control strategy. The article [21] developed a secure reinforcement learning framework that combines model-free learning with model-based safety supervision applied to active suspension systems. A model-free reinforcement learning algorithm was proposed based on a deterministic policy gradient and neural network approximation in [22], which learned the state feedback controller from the sampled data of the suspension system. Liu Ming et al. [23] used the deep deterministic policy gradient algorithm (DDPG) as the control strategy of a semi-active suspension system, which adaptively updates the learned reward value based on the control strategy based on the learned reward values and achieved good control results. The above results fully demonstrate that the use of reinforcement learning algorithms as the control strategy for a semi-active suspension system does not require a strict theoretical definition and accurate mathematical modeling, which can greatly reduce the amount of computation and simplify the complexity of the algorithm. It cannot be ignored that the performance of the vehicle suspension is related to vehicle driving comfort and driving safety, and the performance indicators should be different when facing different road conditions. However, existing successful designs do not sufficiently consider the integration of road information with vehicle driving status to guide the design of relevant strategies.

Motivated by integrating the road condition changes with the reward function, a PPO-based semi-active control strategy is proposed to improve the ride comfort and driving safety. The main contributions are listed as follows:

(1) The PPO algorithm was used as the control strategy of the semi-active suspension, and the body acceleration, suspension deflection, and dynamic tire load were selected as the state space to set the reward function of the PPO algorithm to optimize the performance of the semi-active suspension system;

(2) Combining the road variation with the reward function made it possible to improve the suspension performance by dynamically adjusting the weight matrix of the reward function for different levels of roads;

(3) The proposed control strategy was proven to be effective in improving the performance of the suspension by conducting simulations under different vehicle speeds, vehicle models, and road conditions.

The rest of the paper is organized as follows: In Section 2, we give a systematic modeling of road disturbances and a two-degree-of-freedom (DOF) quarter semi-active suspension system model. Details of the implementation of a semi-active suspension system based on the proximal policy optimization (PPO) algorithm are given in the Section 3. The simulation experiments and results are given in Section 4. Finally, Section 5 gives a conclusion and future work.

2. System Modeling of Road Disturbance and Vehicle Suspension

2.1. Road Disturbance Model

The suspension performance of a vehicle is affected by many factors, including the roughness of the road. According to the international ISO 8608 standard [24], special index data $G_q(n)$ are used to describe the characteristics of the road, and the establishment of a frequency domain road model is mainly based on the actual measurement of the road information, including road unevenness. The road roughness can be divided into eight classes, as shown in Table 1, which specifies the upper and lower limits and average values of the corresponding road spectrum for each road class.

The expression for road power spectral density $G_q(n)$ is as follows:

$$G_q(n) = G_q(n_0)\left(\frac{n}{n_0}\right)^{-\omega}, \quad (1)$$

where $G_q(n)$ is the road unevenness coefficient, which is the road power spectral density value at the reference spatial frequency. n is the spatial frequency, which is the reciprocal of the wavelength; n_0 is the reference spatial frequency, which is generally selected as 0.1 m^{-1}; ω is the frequency index.

Table 1. Eight-level classification of road disturbance roughness.

Road Level	Road Unevenness Coefficient $G_q(n_0)/10^{-6}$ m^3 $n_0 = 0.1$ m^{-1}		
	Lower Limit	Geometric Mean	Upper Limit
A	8	16	32
B	32	64	128
C	128	256	512
D	512	1024	2048
E	2048	4096	8192
F	8192	16,384	32,768
G	32,768	65,536	131,072
H	131,072	262,144	524,288

The corresponding time domain model of filtered white noise road unevenness is as follows [25]:

$$\dot{z}_r = -2\pi f_0 v z_r + 2\pi n_0 \sqrt{G_q(n_0) v} W, \tag{2}$$

where z_r is the road input displacement, v is the vehicle speed, f_0 is the lower cutoff frequency, which is generally chosen as 0.01 m^{-1}, and W is Gaussian white noise with a mean of zero and a sampling frequency of 10 Hz. Different levels of road will cause different amplitudes of vehicle vibration, which affect the suspension performance. In this paper, the PPO algorithm was used as the control strategy to improve the suspension performance for the change of suspension performance caused by road jitter.

2.2. Two-DOF Quarter Semi-Active Vehicle Suspension

The suspension system is affected by many factors, and correspondingly, there are models of suspension systems with complex and diverse degrees of freedom. The two-degree-of-freedom suspension is the most refined structure in the field of automotive suspensions, which can meet the basic requirements of vibration control studies for different levels of roads by intuitively responding to the vibration of a single tire. In this paper, by studying the control strategy of the 2-DOF suspension model, the advantages and disadvantages of the control method can be roughly judged, and the system structure and calculation are relatively simple, which is the basis for the study of the control system of the full vehicle [26]. The 2-DOF quarter semi-active suspension system model is shown in Figure 1.

m_s is the sprung mass; m_u is the unsprung mass; k_s is the sprung stiffness; k_t is the tire stiffness; c_s is the damping coefficient of the suspension; U is the damping force of the semi-active suspension; z_s is the sprung mass displacement; z_u is the unsprung mass displacement; z_r is the road input displacement.

According to the system model in Figure 1 and Newton's second law, the following mechanical model equations of motion can be obtained:

$$\begin{cases} \ddot{z}_s = -\frac{k_s}{m_s}(z_s - z_u) - \frac{c_s}{m_s}(\dot{z}_s - \dot{z}_u) + \frac{U}{m_s}, \\ \ddot{z}_u = -\frac{k_s}{m_u}(z_u - z_s) + \frac{c_s}{m_u}(\dot{z}_s - \dot{z}_u) - \frac{k_t}{m_u}(z_u - z_r) - \frac{U}{m_u}, \end{cases} \tag{3}$$

where $\ddot{z}_s$ is the body acceleration as one of the suspension performance parameter, and the other two choose the suspension deflection $z_s - z_u$ and tire dynamic load $k_t(z_u - z_r)$, through dynamic adjustment of the three scale coefficients for multi-control objective optimization, so that the three control objectives can have a certain optimization effect at the same time.

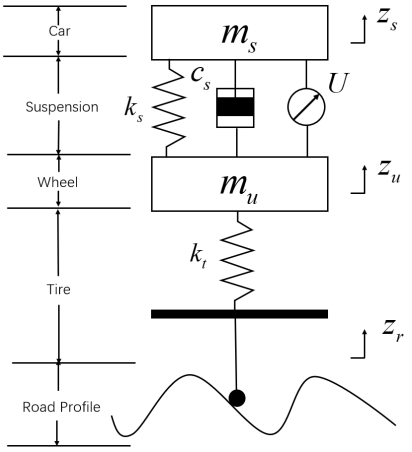

Figure 1. The 2-DOF quarter semi-active suspension system model.

3. Semi-Active Suspension Control Based on the PPO Algorithm

In this paper, the PPO algorithm was combined with a semi-active suspension, and the overall process is that the road information is collected by sensors and input into the suspension system model, while the control performance index of the current moment t is calculated and input into the actor network of the PPO algorithm as the state value; then, the corresponding actor value is selected as the output according to the probability density function in the policy network; a series of trajectories $\tau_i = \{s_i, a_i, r_i, s_{i+1}\}$ is stored in the memory space by repeatedly interacting with the environment; the value network is updated with importance sampling; the control strategy is continuously optimized according to the reward value obtained, and so on, until the control performance is no longer better and convergence is reached. The flowchart of the semi-active suspension control structure based on the PPO algorithm is shown in Figure 2.

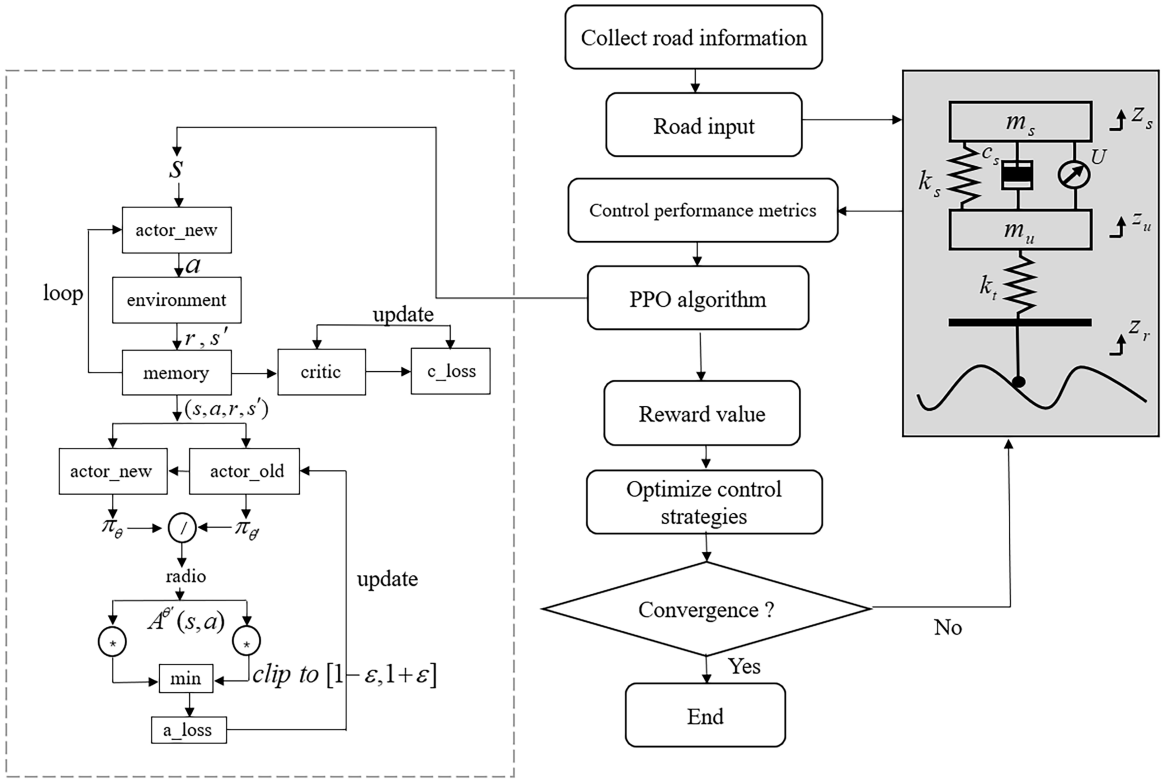

Figure 2. The control structure of semi-active suspension based on PPO.

3.1. State Space and Action Space

According to the above process, the road information is input into the suspension system by using sensors to obtain the corresponding suspension response performances. In this paper, the response indicators φ of the suspension system were selected as the body acceleration, suspension deflection, and tire dynamic load, and the three performance indicators were used as the state space S input for reinforcement learning; the action space A output is the damping force of the suspension system; considering the practical application of the suspension system, the range of the damping force was set to −500 N to 500 N. Figure 3 shows a change in the damping force of the semi-active suspension. The resulting response performance φ is input as the state space into the policy network of the PPO algorithm.

$$\varphi = \begin{pmatrix} \ddot{z}_s \\ z_s - z_u \\ k_t(z_u - z_r) \end{pmatrix}. \tag{4}$$

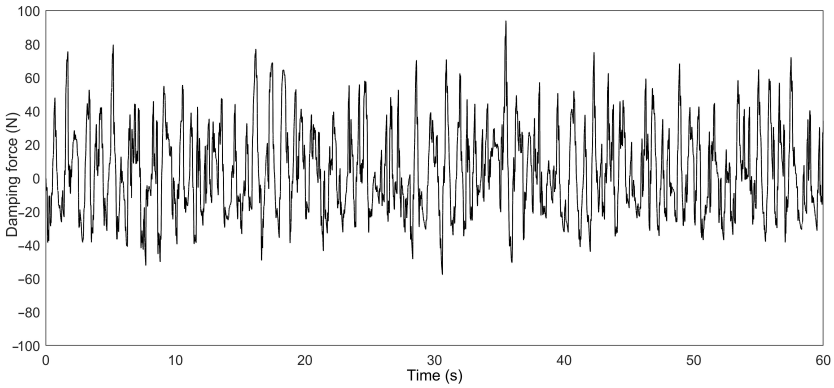

Figure 3. The changing of the damping force of the semi-active suspension.

3.2. PPO Algorithm Network Model

The proximal policy optimization (PPO) algorithm is an actor–critic-based reinforcement learning algorithm, which combines a value-based approach and a policy-based approach [27]. The architecture of the algorithm consists of two parts: the actor network part is based on the policy approach, which is used to optimize the policy model $\pi(s, a)$ by generating actions and interacting with the environment; the critic network part is based on the value approach, which is used to determine the merit of an action and select the next action accordingly to optimize the value function $Q(s, a)$. The network structure of the PPO algorithm is shown in Figure 4.

It contains four neural networks [28], which are the policy network, the target policy network, the value network, and the target value network. The network architectures of the policy network and the value network are the same as those of their corresponding target networks, and there are only differences in the network parameters. In the architecture of this control algorithm, the actor network obtains the control data based on the output state information of the suspension system, so as to obtain the action information damping force to optimize the entire algorithm. By adjusting different speeds and road bumps, the car's body acceleration and suspension deflection at the current moment are affected, resulting in the next new state.

After a period of interaction between the agent and the environment, a series of trajectory data is generated, and the network parameters are updated by sampling the data stored in the memory network using a policy gradient. In the policy gradient (PG) [29] method, the goal is to learn a parameterized policy that determines a unique action or obtains a corresponding action probability density function directly from a selection in the state space. In the learning process, the strategy is optimized by calculating an estimate of

the policy gradient, and the methods used are gradient up or gradient down depending on the expectation, respectively. The policy gradient function is shown in Equation (5):

$$\nabla \bar{R}_\theta = E_{(s_t,a_t)\sim\pi_\theta}[R(s_t,a_t)\nabla \log_{\pi_\theta}(s_t,a_t)], \tag{5}$$

$$\begin{aligned} A^{\pi,\gamma}(s_t,a_t) &= E_{s_{t+1}}[r_t + \gamma V^{\pi,\gamma}(s_{t+1}) - V^{\pi,\gamma}(s_t)] \\ &= E_{s_{t+1}}[Q^{\pi,\gamma}(s_t,a_t) - V^{\pi,\gamma}(s_t)], \end{aligned} \tag{6}$$

where π_θ is a stochastic policy function and $R(s_t,a_t)$ is the reward value of the current state action pair. $V^{\pi,\gamma}(s_t)$ is the state value function of the strategy π at the current moment. $Q^{\pi,\gamma}(s_t,a_t)$ is the action value function of the strategy π at the current moment. $A^{\pi,\gamma}(s_t,a_t)$ is an estimate of the advantage function at the moment t [30], containing a hyperparameter γ, which is usually obtained by estimating the advantage function based on the value of the reward function, and then, the policy gradient is estimated according to the parameter θ so as to optimize the policy with the policy gradient. If $A^{\pi,\gamma}(s_t,a_t) > 0$, this means that the current state action pair (s_t,a_t) is able to obtain a higher reward than the baseline, and conversely, a lower reward is obtained.

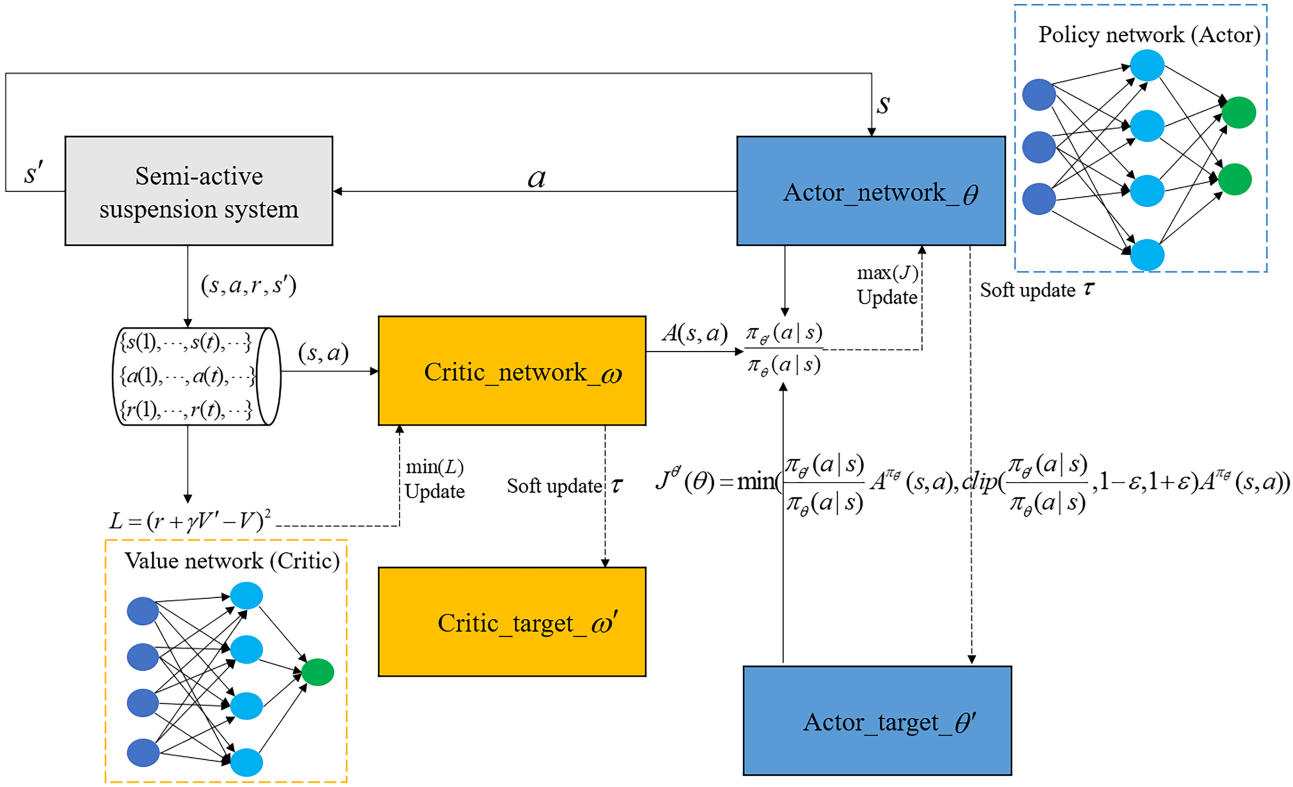

Figure 4. The PPO algorithm's structure.

3.3. Design If the Reward Function with Road Disturbances and Policy Update

The process of policy updating depends on the reward value of the agent's state action pair to judge its goodness. The reward function in this paper was set for a semi-active suspension optimized for multiple objectives, which considered the vehicle acceleration on the human body, as well as the vehicle body more to improve the performance. Therefore, it is only necessary to ensure that the remaining two performance indicators have a certain improvement or some weakening within the acceptable range of the vehicle suspension performance. Considering different reward functions depending on the optimization task, we chose to make it as simple as possible to evaluate the efficiency of reinforcement learning optimization.

Therefore, in this paper, the performance metrics of the passive suspension for different road conditions were directly set up with a reward and punishment mechanism, giving a positive reward value if the performance improves and, conversely, giving a negative reward value. The reward function R was set by summing up the suspension performance φ caused by different road conditions z_r and dynamically adjusting the weight matrix of the suspension performance according to different road conditions. For example, when the road is more rough and the safety requirement of the vehicle is greater, the weight coefficient corresponding to the tire load is increased. Therefore, the reward function is shown as follows:

$$z_r \rightarrow \varphi, \tag{7}$$

$$R = \sum_{t=0}^{T} \varphi^{\mathrm{T}} A \varphi, \tag{8}$$

$$A = \begin{pmatrix} x & & \\ & y & \\ & & z \end{pmatrix}, \tag{9}$$

where A represents the weight matrix of each performance indicator, and the values of x, y, and z are dynamically adjusted according to different roads. For example, it was experimentally measured that when a light vehicle with a speed of 20 m/s is driving on a C-class road, the suspension response performance φ is better when $A = \begin{pmatrix} 10^3 & & \\ & 1.5 & \\ & & 1 \end{pmatrix}$.

When using the reward function and the policy gradient method to update the network, the traditional on-policy policy was used, and the policy that interacts with the environment π and the policy that needs to be updated π' were the same policy, the disadvantage being that the sampling efficiency was low and the data sampled each time could only be used for one policy update. Therefore, the PPO algorithm introduces another policy q to interact with the environment, and the empirical samples obtained from the interaction are placed in memory and can be used to update the policy p several times. According to Equation (10), it was proven that the purpose can be achieved by multiplying only one weighting factor $\frac{p(x)}{q(x)}$ on the basis of the original.

$$E_{x\sim p}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = E_{x\sim q}[f(x)\frac{p(x)}{q(x)}], \tag{10}$$

$$J^{\theta'}(\theta) = E_{(s_t,a_t)\sim\pi_{\theta'}}[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}A^{\theta'}(s_t,a_t)]. \tag{11}$$

In Equation (11), $J^{\theta'}(\theta)$ is the target function, p and q correspond to the specific policy distributions π_θ and $\pi_{\theta'}$, respectively, and $A^{\theta'}$ is the advantage function of the t moment, that is the gap between the score and the baseline obtained by the current strategy and the environment interaction. PPO uses importance sampling to change the previous on-policy policy into an off-policy policy [31], making the sampled data reusable and improving the learning efficiency. This was done by using two policy networks, one π_θ to interact with the environment and the other $\pi_{\theta'}$ to update the network.

$\hat{E}_t[\cdots]$ represents the expectation of a batch of samples, and the obtained expectation is approximately equal when the policy distribution of π_θ and $\pi_{\theta'}$ is similar; otherwise, it will produce a large error. Therefore, PPO is constrained by a first-order optimization, converting Equation (11) to Equation (12) as shown in:

$$J^{\theta'}(\theta) = \min(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}A^{\pi_{\theta'}}(s,a), clip(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}, 1-\varepsilon, 1+\varepsilon)A^{\pi_{\theta'}}(s,a)), \tag{12}$$

where ε is the hyperparameter of the clip function, which is used to limit the magnitude of the update; the second half uses the clip function $clip(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}, 1-\varepsilon, 1+\varepsilon)A^{\pi_{\theta'}}(s,a)$ to limit π_θ and $\pi_{\theta'}$ from being too different. This clip idea is that when $A > 0$, this means that the current action produces a return estimate that is greater than the expected return of the baseline, so we update the strategy π so that the probability of that action appearing is as high as possible, but to add a limit to this, that is it cannot be $1+\varepsilon$-times higher than the original strategy. Similarly, if $A < 0$, this means that the estimated return of the current action is less than the expected return of the baseline, so we made the probability of the action occurring, as low as possible, but not less than $1-\varepsilon$-times the original strategy.

3.4. Semi-Active Suspension Control Based on the PPO Architecture

In view of the above ideas, the overall process of summarizing the algorithm is shown in Table 2.

Table 2. Proximal policy optimization algorithm-semi-active suspension.

Procedure: PPO-semi-active suspension

The on-board sensor is used to collect road information in time for the semi-active suspension system to obtain the suspension response performance φ.
Randomly initialize the actor network and critic network with weights θ and ω.
Initialize the actor target network and critic target network: $\theta' \leftarrow \theta$, $\omega' \leftarrow \omega$.
for $k = 0, 1, 2, \cdots$ **do**
Generate a set of trajectories $\tau_i = \{s_i, a_i, r_i, s_{i+1}\}$ by running policy π_{θ_k} in the environment.
Sample the data information in the memory network, and calculate the quality of the corresponding state action pair (s_i, a_i) according to the reward function R.
Compute the advantage function $A^{\theta'}(s_t, a_t) = r(s_t, a_t) + \gamma V^{\omega_k}(s_{t+1}) - V^{\omega_k}(s_t)$ based on the current value function V_{ω_k}.
Update the policy parameter by maximizing the PPO-clip objective:

$$\theta_{k+1} = \arg\max_\theta \frac{1}{|\tau_k|T} \sum_{t=0}^{T} \min\left(\frac{\pi_{\theta_k}(a_t|s_t)}{\pi_\theta(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), clip\left(\frac{\pi_{\theta_k}(a_t|s_t)}{\pi_\theta(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\pi_{\theta_k}}(s_t, a_t)\right).$$

Update the value parameter by minimizing the advantage function:

$$\omega_{k+1} = \arg\min_\omega \frac{1}{|\tau_k|T} \sum_{t=0}^{T} \alpha\left(r(s_t, a_t) + \gamma V^{\omega_k}(s_{t+1}) - V^{\omega_k}(s_t)\right)^2.$$

Soft update the target network parameters:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta',$$
$$\omega' \leftarrow \tau\omega + (1-\tau)\omega'.$$

end for

4. Simulation Results and Analysis

In this paper, MATLAB/SIMULINK was used to build a 1/4 semi-active suspension system model, in which the suspension parameters are shown in Table 3. The actor network and critic network of the PPO algorithm were neural networks with five and six hidden layers. The learning rate of the actor network α_A and the critic network α_C was set to 0.001 and 0.0001, respectively. The discount factor γ was set to 0.98. The clipparameter ε was set to 0.2. The GAE parameter λ was set to 0.9. The batch size was set to 64. The simulation time was set to 60 s. The specific hyperparameters were set as shown in Table 4.

Table 3. Parameters for the quarter semi-active suspension.

Parameters	Symbols	Unit	Light	Heavy
Sprung mass	m_s	(kg)	167.8	1465.4
Unsprung mass	m_u	(kg)	22.3	122.3
Sprung stiffness	k_s	(N/m)	11,530	11,530
Tire stiffness	k_t	(N/m)	115,300	315,300
Damping coefficient of suspension	c_s	(N*m/s)	278.4	4670.4

Table 4. Hyperparameters for the proximal policy optimization structure.

Hyperparameters	Symbols	Values
Actor learning rate	α_A	0.0002
Critic learning rate	α_C	0.001
Discount factor	γ	0.98
Clip parameter	ε	0.2
GAE parameter	λ	0.9
Batch size		64
Time steps per batch		600

4.1. Simulation Experiments for Different Road Levels for a Light Vehicle

In order to verify the control performance and robustness of the semi-active suspension system based on the PPO algorithm, simulation comparison experiments were conducted for the passive suspension system, the semi-active suspension system based on fuzzy control, and the semi-active suspension system based on the PPO algorithm under the same conditions. Specifically, a vehicle speed of 20 m/s was simulated on the C-, D-, and E-class roads, respectively, and the road power spectrum is shown in Figure 5.

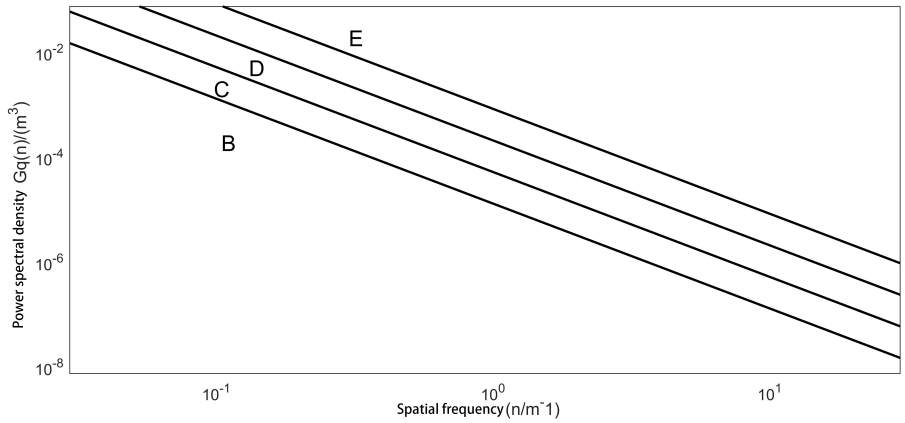

Figure 5. The road grade distribution.

The simulation experimental results of the body acceleration, suspension deflection, and dynamic tire load on C-class and D-class roads at a vehicle speed of 20 m/s are shown in Figures 6 and 7. For the ease of comparison, the percentages of the root mean square (RMS) values for the body acceleration, suspension deflection, and tire dynamic load of the passive suspension, the semi-active suspension system based on fuzzy control, and the PPO algorithm under the same conditions are given in Table 5, and the corresponding root-mean-squared error (RMSE) values are given in Table 6.

The simulation results showed that the body accelerations of the PPO-based semi-active suspension system and the fuzzy-based system were reduced by 59.1% and 26.3%, the suspension deflections were reduced by 45.5% and 27%, and the dynamic tire loads were reduced by 28% and 13.3%, compared with the passive suspension system under the D-class road.

Table 5. The percentages of RMS values of the two semi-active suspensions for a light vehicle under the same conditions.

Road Level	Body Acceleration		Suspension Deflection		Tire Dynamic Load	
	Fuzzy	PPO	Fuzzy	PPO	Fuzzy	PPO
C	26.4%	61.8%	6.1%	47.8%	−18%	28.8%
D	26.3%	59.1%	27%	45.5%	13.3%	28%
E	26.4%	59.3%	27.1%	46.5%	13.3%	28.8%

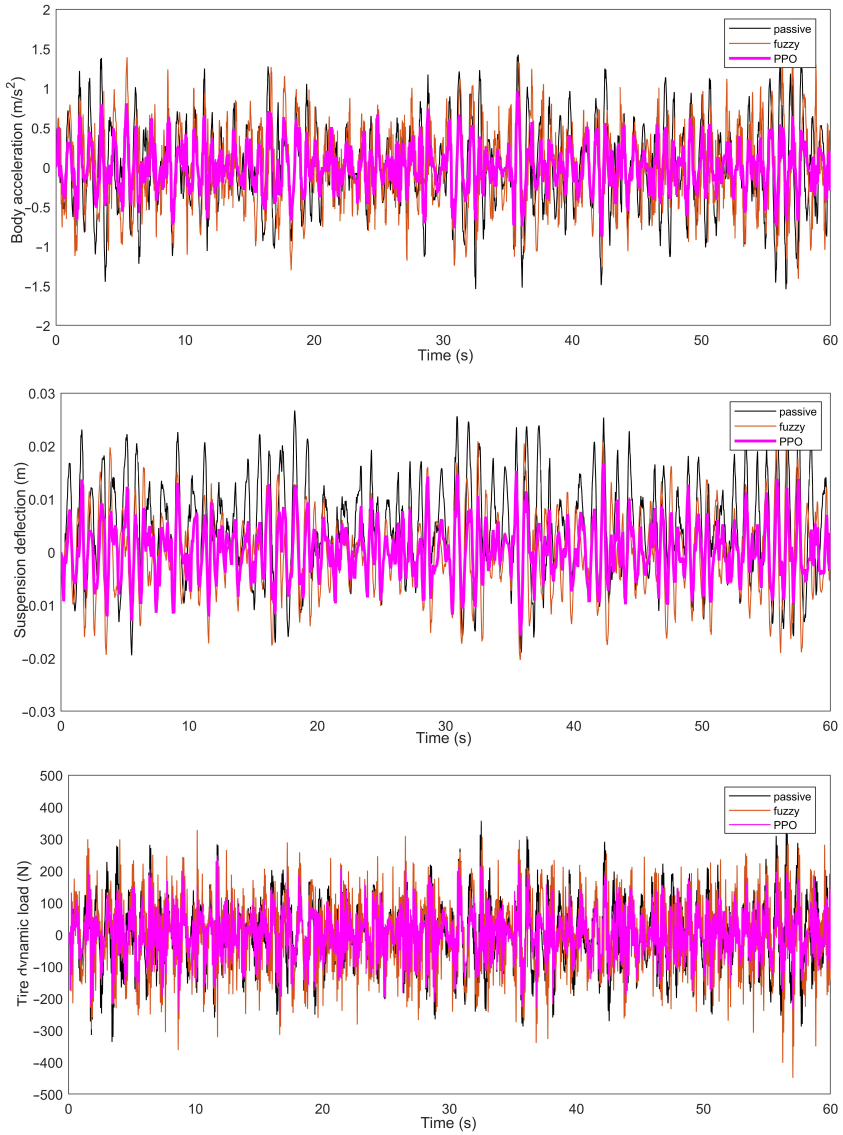

Figure 6. Simulation results of fuzzy, PPO, and passive suspension systems for a light vehicle on a C-class road.

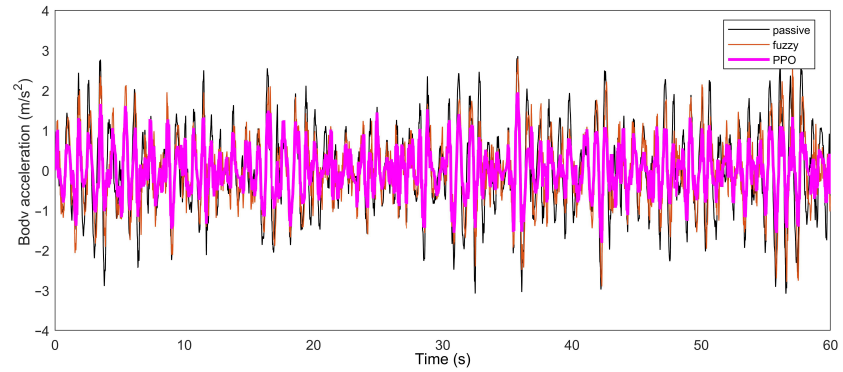

Figure 7. *Cont.*

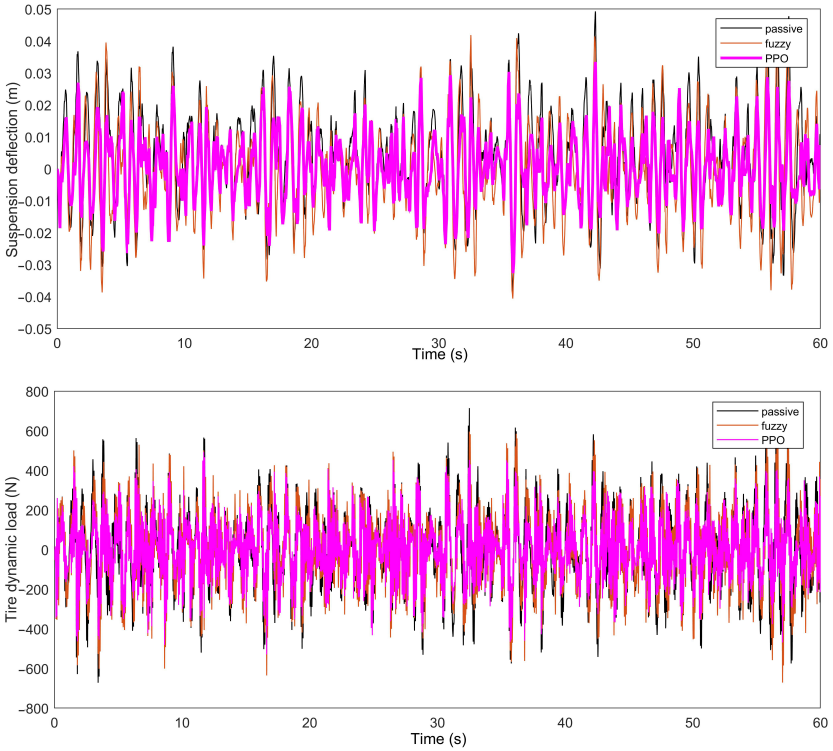

Figure 7. Simulation results of fuzzy, PPO, and passive suspension systems for a light vehicle on a D-class road.

Table 6. The RMSE values of the two semi-active suspensions for a light vehicle under the same conditions.

Road Level	Body Acceleration			Suspension Deflection			Tire Dynamic Load		
	Passive	Fuzzy	PPO	Passive	Fuzzy	PPO	Passive	Fuzzy	PPO
C	0.5896	0.4735	0.3492	0.1275	0.06218	0.05513	12.71	9.624	11.93
D	0.8338	0.6844	0.6614	0.1424	0.08794	0.07839	18.77	16.86	13.61
E	1.179	1.058	0.9787	0.1671	0.1244	0.1109	25.45	24.23	19.25

4.2. Simulation Experiments for Different Road Levels for a Heavy Vehicle

In order to test the adaptability of the control strategy proposed in this paper to different vehicle models, simulation experiments were conducted for a heavy vehicle on B-, C-, and D-class roads by changing the vehicle speed to 15 m/s according to the actual situation, and the suspension parameters and road power spectrum are shown in Table 3 and Figure 5, respectively. Figure 8 shows the simulation results of the body acceleration on B- and C-class roads. The percentages of the RMS values of the body acceleration for the passive suspension and the semi-active suspension system based on the fuzzy and PPO algorithm under the same conditions are given in Table 7, and the corresponding RMSE values are given in Table 8. The experimental results showed that the PPO-based semi-active suspension system still had good performance under different vehicle models, vehicle speeds, and road conditions.

Table 7. The percentages of the RMS values of the two semi-active suspensions for a heavy vehicle under the same conditions.

Percentage of Body Acceleration	B	C	D
Fuzzy	18.1%	3.8%	5%
PPO	44.4%	31.6%	10%

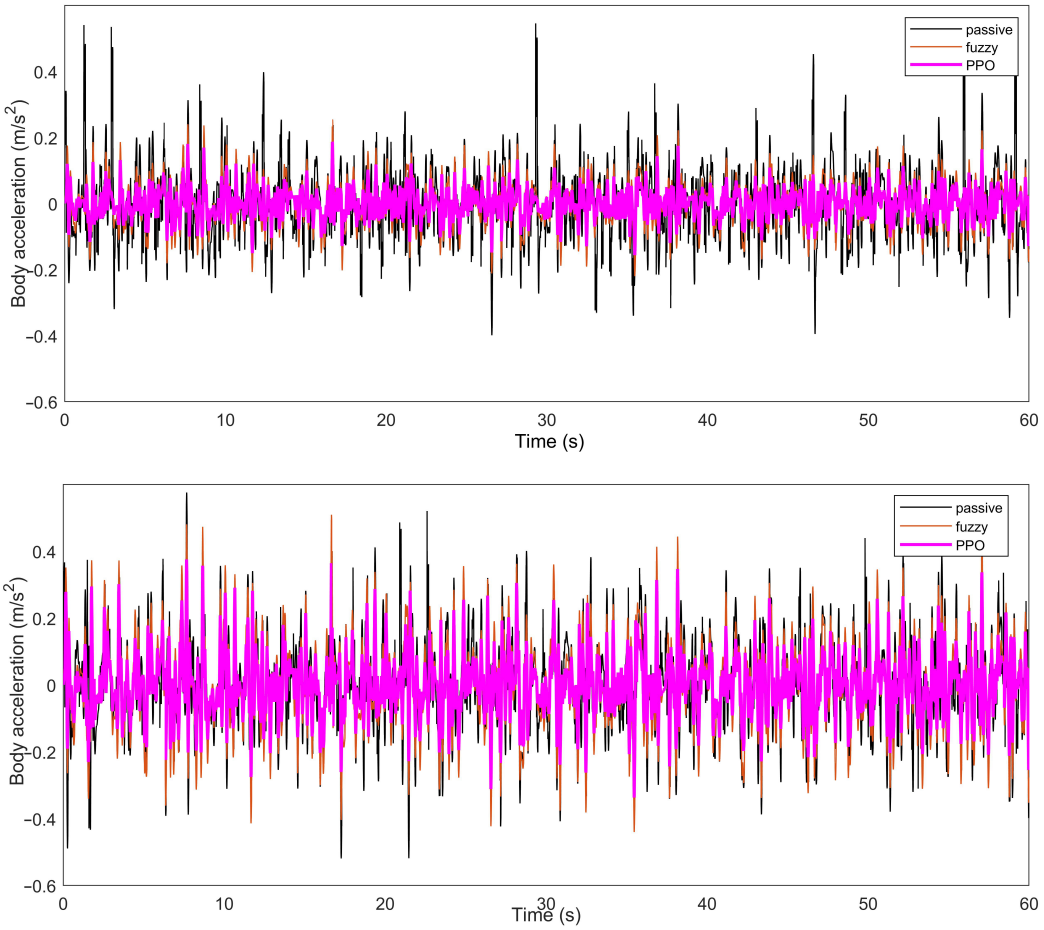

Figure 8. Body acceleration of fuzzy, PPO, and passive suspension systems for a heavy vehicle on class-B and -C roads.

Table 8. The RMSE values of the two semi-active suspensions for a heavy vehicle under the same conditions.

Body Acceleration	B	C	D
Passive	0.218	0.3083	0.436
Fuzzy	0.1973	0.3023	0.425
PPO	0.1625	0.255	0.4147

4.3. Simulation Experiments for a Continuously Changing Road Level

In order to verify the applicability of the proposed control strategy to the case of the road level changing randomly, simulation experiments were conducted for a light vehicle at a speed of 20 m/s with continuous changes in the road grade, and the road surface changes and body acceleration are shown in Figures 9 and 10. As shown in Figure 9, the road changed from Class B to Class C, then to Class D, and finally, back to Class C. The duration of each section was 10 s, and the total simulation time was 40 s. According to the simulation experiment results in Figure 10, it was proven that the semi-active suspension system based on PPO still had good performance under the continuously changing roads, and the body acceleration value was reduced by 46.93%.

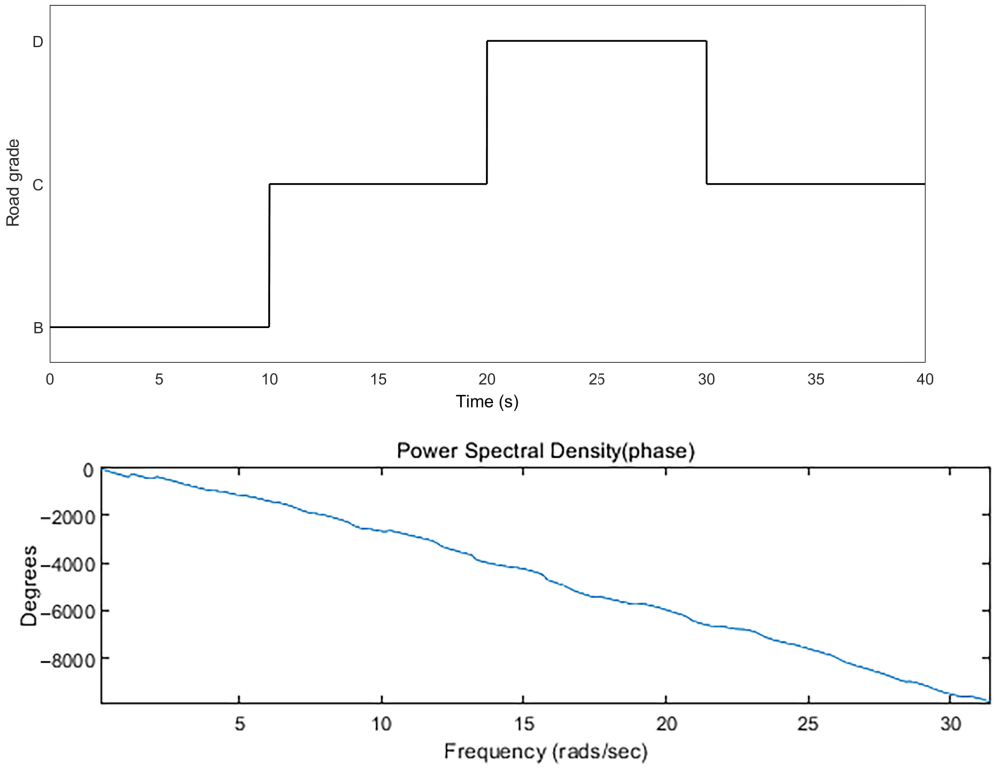

Figure 9. Continuously changing road.

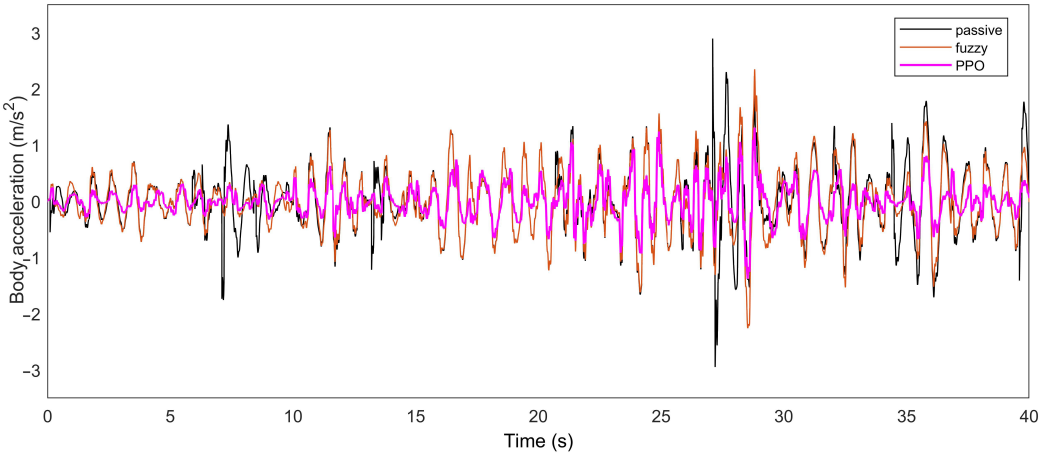

Figure 10. Body acceleration under continuously a changing road of PPO, passive, and fuzzy suspension systems for a light vehicle.

4.4. Number of Iterations and Computation Time

This paper used the PPO algorithm for training. The max episodes was set to 10,000. The max steps per episode was set to 600. The sampling time was 0.1 s. The time per iteration was 60 s. The mini batch size was set to 64. The convergence process of the algorithm is shown in Figure 11, which tended to converge at 90 episodes according to the figure, with a total time of about 1.2 h, and it took a short time from the beginning of the algorithm to search for good behavior to the final convergence, which was faster than the number of iterations of the particle swarm optimization (PSO) algorithm in the paper [32].

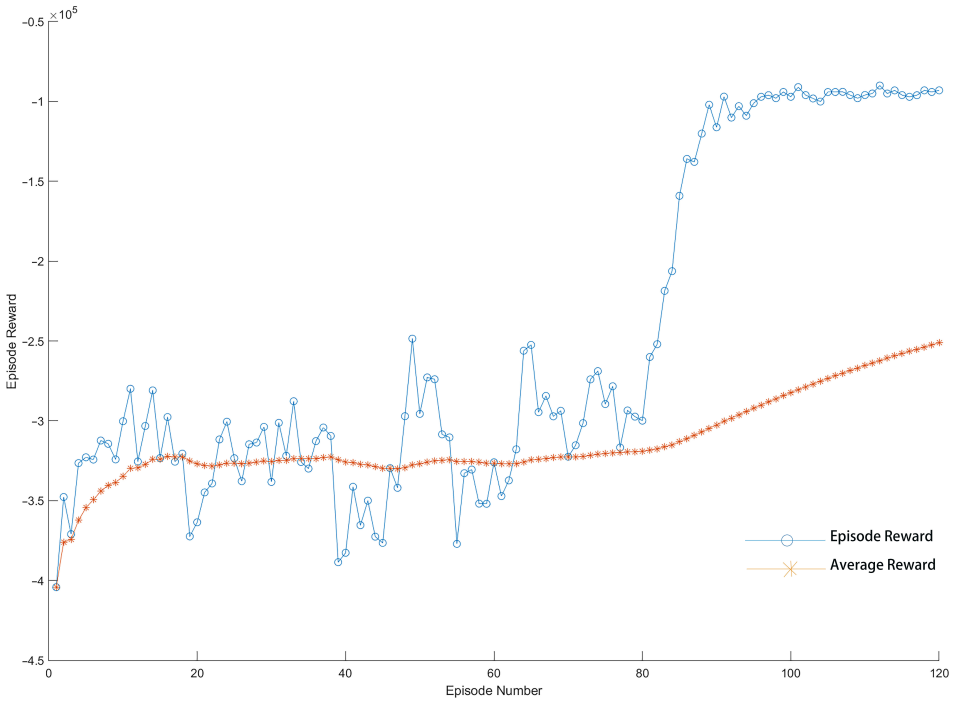

Figure 11. Number of iterations.

5. Conclusions and Future Work

In this paper, a semi-active suspension control strategy based on the Proximal Policy Optimization (PPO) algorithm was proposed. Firstly, the road unevenness information was collected in real time by the sensor, and the three performance indicators of the vehicle were obtained as the state input for reinforcement learning by establishing a 2-DOF quarter suspension semi-system model, then the control strategy was continuously optimized according to the corresponding reward and punishment functions and the prints of the weight matrix under different road conditions until the convergence state was reached. The simulation experiments showed that the proposed control strategy could improve the vehicle performance better than the passive suspension and fuzzy control.

However, the traditional reward function is not applicable in the face of complex and frequently changing roads, and the applicability of the control strategy involved to the full-vehicle suspension requires further exploration and research.

Author Contributions: Conceptualization, S.-Y.H.; methodology, S.-Y.H.; writing—original draft preparation, T.L.; writing—review and editing, S.-Y.H. and T.L.; supervision, S.-Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province for Key Project under Grant ZR2020KF006, the National Natural Science Foundation of China under Grants 61903156 and 61873324, "New 20 Rules for University" Program of Jinan City under Grant No. 2021GXRC077, the Natural Science Foundation of Shandong Province under Grant ZR2019MF040, the University Innovation Team Project of Jinan under Grant 2019GXRC015, the Higher Educational Science and Technology Program of Jinan City under Grant 2020GXRC057, and the State Scholarship Fund of the China Scholarship Council.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Du, H.-P.; Sze, K.Y.; Lam, J. Semi-active H-infinity control of vehicle suspension with magnetorheological dampers. *J. Sound Vib.* **2005**, *283*, 981–996. [CrossRef]
2. Rao, L.; Narayanan, S. Sky-hook control of nonlinear quarter car model traversing rough road matching performance of LQR control. *J. Sound Vib.* **2009**, *323*, 515–529.
3. Zhao, Y.; Sun, W.; Gao, H. Robust control synthesis for seat suspension systems with actuator saturation and time-varying input delay. *J. Sound Vib.* **2010**, *329*, 4335–4353. [CrossRef]
4. Pepe, G.A. VFC—Variational feedback controller and its application to semi-active suspensions. *Mech. Syst. Signal Process.* **2016**, *76*, 72–92. [CrossRef]
5. Wu, W.; Chen, X.; Shan, Y. Analysis and experiment of a vibration isolator using a novel magnetic spring with negative stiffness. *J. Sound Vib.* **2014**, *333*, 2958–2970. [CrossRef]
6. Zhang, M.-H.; Jing, X.-J. A bioinspired dynamics-based adaptive fuzzy SMC method for half-car active suspension systems with input dead zones and saturations. *IEEE Trans. Cybern.* **2021**, *51*, 1743–1755. [CrossRef]
7. Bui, Q.-D.; Nguyen, Q.H. A new approach for dynamic modeling of magnetorheological dampers based on quasi-static model and hysteresis multiplication factor. In Proceedings of the IFToMM Asian Conference on Mechanism and Machine Science, Hanoi, Vietnam, 15–18 December 2021; pp. 733–743.
8. Bui, Q.; Hoang, L.; Mai, D.; Nguyen, Q.H. Design and testing of a new shear-mode magnetorheological damper with self-power component for front-loaded washing machines. In Proceedings of the 2nd Annual International Conference on Material, Machines and Methods for Sustainable Development (MMMS2020), Trang, Vietnam, 12–15 November 2020; pp. 860–866.
9. Phu, D.X.; Mien, V. Robust control for vibration control systems with dead-zone band and time delay under severe disturbance using adaptive fuzzy neural network. *J. Frankl. Inst.* **2020**, *357*, 12281–12307. [CrossRef]
10. Phu, D.X.; Mien, V.; Tu, P.; Nguyen, N.P.; Choi, S.B. A new optimal sliding mode controller with adjustable gains based on Bolza–Meyer criterion for vibration control. *J. Sound Vib.* **2020**, *485*, 115542. [CrossRef]
11. Mnih, V.; Kavukcuoglu, K.; Silver, D. Playing Atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
12. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
13. Schulman, J.; Levine, S.; Moritz, P. Trust region policy optimization. In Proceedings of the 2015 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1889–1897.
14. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
15. Ye, D.; Liu, Z.; Sun, M. Mastering complex control in MOBA games with deep reinforcement learning. *AAAI Conf. Artif. Intell.* **2020**, *34*, 6672–6679. [CrossRef]
16. Schulman, J.; Wolski, F.P. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
17. Le, Z.; Yza, B.; Xin, Z.A. Image captioning via proximal policy optimization. *Image Vis. Comput.* **2021**, *108*, 104126.
18. Sadhukhan, P.; Selmic, R.R. Multi-agent formation control with obstacle avoidance using proximal policy optimization. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 2694–2699.
19. Tognetti, S.; Savaresi, S.M.; Spelta, C. Batch reinforcement learning for semi-active suspension control. In Proceedings of the 2009 IEEE Control Applications, (CCA) and Intelligent Control, St. Petersburg, Russia 8–10 July 2009; pp. 582–587.
20. Bucak, I.O.; Oez, H.R. Vibration control of a nonlinear quarter-car active suspension system by reinforcement learning. *Int. J. Syst. Sci.* **2012**, *43*, 1177–1190. [CrossRef]
21. Li, Z.; Chu, T.; Kalabic, U. Dynamics-enabled safe deep reinforcement learning: Case study on active suspension control. In Proceedings of the IEEE Conference on Control Technology and Applications (CCTA), Hong Kong, China, 19–21 August 2019; Volume 19, pp. 585–591.
22. Zhao, F.; Jiang, P.; You, K.; Song, S.; Zhang, W.; Tong, L. Setpoint tracking for the suspension system of medium-speed maglev trains via reinforcement learning. In Proceedings of the 2019 IEEE 15th International Conference on Control and Automation (ICCA), Edinburgh, UK, 16–19 July 2019; pp. 1620–1625.
23. Liu, M.; Li, Y.; Rong, X. Semi-active suspension control based on deep reinforcement learning. *IEEE Access* **2020**, *8*, 9978–9986.
24. Technical Committee ISO/TC, Mechanical Vibration, Shock. Subcommittee SC2 Measurement, Mechanical vibration-road surface profiles-reporting of measured data. *Int. Organ. Stand.* **1995**, *8608*. Available online: https://kns.cnki.net/kcms/detail/detail.aspx?FileName=SCSF00013909&DbName=SCSF (accessed on 1 January 2006).
25. Zhang, Y.; Zhang, J. Numerical simulation of stochastic road process using white noise filtration. *Mech. Syst. Signal Process.* **2006**, *20*, 363–372.
26. Nawathe, P.R.; Shire, H.; Wable, V. Simulation of passive suspension system for improving ride comfort of vehicle. *Int. J. Manag. Technol. Eng.* **2018**, *8*, 401–411.
27. Du, W.; Ding, S. A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications. *Artif. Intell. Rev.* **2021**, *54*, 3215–3238. [CrossRef]
28. Obando-Ceron, J.S.; Castro, P.S. Revisiting rainbow: Promoting more insightful and inclusive deep reinforcement learning research. In Proceedings of the 2021 International Conference on Machine Learning (ICML), Online, 18–24 July 2021; pp. 1373–1383.

29. Yoo, H.; Kim, B.; Kim, J.W. Reinforcement learning based optimal control of batch processes using Monte-Carlo deep deterministic policy gradient with phase segmentation. *Comput. Chem. Eng.* **2021**, *144*, 107133. [CrossRef]
30. Zimmer, M.; Weng, P. Exploiting the sign of the advantage function to learn deterministic policies in continuous domains. In Proceedings of the 2019 International Joint Conferences on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 4496–4502.
31. Zhang, H.; Bai, S.; Lan, X. Hindsight trust region policy optimization. In Proceedings of the 2019 International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 19–27 August 2021; pp. 3335–3341.
32. Qu, Q.; Qi, M.; Gong, R. An improved enhanced fireworks algorithm based on adaptive explosion amplitude and levy flight. *Eng. Lett.* **2020**, *28*, 1348–1357.