World Electric Vehicle Journal

MDPI

Article

Deep Reinforcement Learning Algorithm Based on Fusion Optimization for Fuel Cell Gas Supply System Control

Hongyan Yuan, Zhendong Sun, Yujie Wang and Zonghai Chen *

Department of Automation, University of Science and Technology of China, Hefei 230027, China
* Correspondence: chenzh@ustc.edu.cn

Abstract: In a proton exchange membrane fuel cell (PEMFC) system, the flow of air and hydrogen is the main factor affecting the output characteristics of the PEMFC, and there is a coordination problem in the flow control of both. To ensure real-time gas supply in the fuel cell and improve the output power and economic benefits of the system, a deep reinforcement learning controller with continuous state based on fusion optimization (FO-DDPG) and a control optimization strategy based on net power optimization are proposed in this paper, and the effects of whether the two gas controls are decoupled or not are compared. The experimental results show that the undecoupled FO-DDPG algorithm has a faster dynamic response and more stable static performance compared to the fuzzy PID, DQN, traditional DRL algorithm, and decoupled controllers, demonstrated by a dynamic response time of 0.15 s, an overshoot of less than 5%, and a steady-state error of 0.00003.

Keywords: proton exchange membrane fuel cell (PEMFC); gas supply system control; deep reinforcement learning; fusion optimization

1. Introduction

The fuel cell (FC) is considered to be an alternative to traditional fossil energy sources such as coal and petroleum in the 21st century. The FC provides an effective approach to solve the shortage of petroleum energy and environmental pollution and is also related to future energy security and long-term development. According to the different electrolytes, fuel cells can be divided into alkaline fuel cell (AFC), phosphoric acid fuel cell (PAFC), molten carbonate fuel cell (MCFC), solid oxide fuel cell (SOFC), and proton exchange membrane fuel cell (PEMFC). Among them, the PEMFC is the most widely used.

Compared with fossil energy sources, the PEMFC technology has the following advantages [1]: high energy conversion efficiency, with an expected efficiency of more than 80% [2]; low or even no emissions can be achieved; low noise, high reliability and easy maintenance during operation; and the power generation efficiency is less affected by the load. Therefore, the PEMFC is widely used in many fields such as aerospace, military, electric vehicle, and distributed electricity applications [3]. However, fuel cells also have many shortcomings: the fuel cell reaction requires the action of catalysts and has high requirements for catalysts; the use of hydrogen energy as fuel, the storage and transportation of hydrogen energy have safety and stability problems; the output characteristics are soft and need to improve the control performance, reduce the impact on the battery due to load changes and the external environment, and improve the battery life [4,5].

The safe and efficient operation of the fuel cell cannot be achieved without the operation of its auxiliary system. The auxiliary system regulates the inlet, exhaust, temperature, and humidity of the stack in real-time according to the load demand power to ensure that the stack always operates under the ideal working environment. A complete fuel cell system (FCS) includes a fuel cell stack, gas supply system, temperature control system, humidity control system and energy management system. Among them, the gas supply system includes oxygen supply and hydrogen supply to provide reactants for the

check for updates

Citation: Yuan, H.; Sun, Z.; Wang, Y.; Chen, Z. Deep Reinforcement Learning Algorithm Based on Fusion Optimization for Fuel Cell Gas Supply System Control. *World Electr. Veh. J.* **2023**, *14*, 50. https://doi.org/10.3390/wevj14020050

Academic Editor: Michael Fowler

Received: 17 January 2023
Revised: 1 February 2023
Accepted: 7 February 2023
Published: 10 February 2023

electrochemical reaction of the fuel cell, which is converted into electricity through the electrochemical reaction. At the same time, the degree of oxygen and hydrogen supply in the fuel cell largely affects the output power of the fuel cell [6]. Therefore, effective control of gases in fuel cells is necessary.

The oxygen supply to the fuel cell is provided by an air compressor that pressurizes the outside air and feeds it to the stack. Due to the high latency of air compressors and other equipment, under- or over-supply of oxygen may occur during load current changes. Insufficient oxygen supply can lead to a decrease in output voltage and power, permanently damaging the fuel cell; excessive oxygen supply can lead to an increase in parasitic power, reducing the net system output power [7]. Therefore, the oxygen flow rate needs to be controlled to ensure the safe and efficient operation of the FCS. In many studies, oxygen excess ratio (OER) is commonly used to measure the degree of oxygen quantity availability, as defined in Equation (5).

The control core of the air supply system is the regulation of the flow of air into the power stack. The main current air system control strategies are classified according to the control methods: traditional control, advanced control, and intelligent control algorithms.

Traditional control strategies mainly include feed-forward control and PID control. Feed-forward control establishes the corresponding logic control strategy through off-line testing and empirical knowledge, which has the advantages of simple design and small calculation, but the control effect is single and relatively poor. PID methods are widely used in industrial process control because of their simple structure and high reliability, but they are prone to overshoot and jitter in some cases. In Ref. [8], feed-forward control and PID control were used for PEMFCs, and the coordinated operation of different control methods was simulated to verify the feasibility of the control strategy. Meanwhile, in order to improve the control accuracy and robustness of traditional PID, researchers have successively proposed many improved PID algorithms, such as fuzzy PID and adaptive PID [9]. Ref. [10] uses an adaptive fuzzy PID control method to control the gas supply system, which ensures a stable variation of the gas supply system with the change of load power demand and prevents the instability that may be brought by the gas supply system.

For the nonlinear and time-varying characteristics of the PEMFC gas supply system, advanced control strategies such as adaptive control, robust control, and predictive control have achieved good results. E. S. Kim [11] designed a suitable state feedback controller to improve the response speed and immunity to disturbances of the PEMFC gas supply system. In Ref. [12], a fuzzy generalized predictive controller based on a control-oriented T-S model is designed to control the oxygen excess ratio in the ideal range and effectively suppress the fluctuation caused by the load change.

Intelligent control mainly uses intelligent algorithms such as neural networks, which have strong robustness and self-learning capability in OER control. M Sedighizadeh [13] proposed an adaptive control strategy based on artificial neural networks to control the PEMFC system to improve the dynamic performance of oxygen flow control. ChunHua Li [14] used a fuzzy adaptive recurrent neural network to model the PEMFC air supply system to prevent the occurrence of oxygen scarcity and improve the system response performance. However, the establishment of neural network models requires a large amount of data support and too much computational effort.

The structures of hydrogen supply systems can be generally classified as anode blind-end type and cyclic type. Studies have shown that the cyclic anode structure has a higher system and stack efficiency when the power is above a certain threshold [15], which has been widely used in recent years. In a circulating hydrogen supply system, the supply of hydrogen is performed by a hydrogen tank, which uses a proportional valve to achieve hydrogen injection and a circulation structure such as a circulation pump to achieve hydrogen recycling. Similar to the OER, the hydrogen excess ratio (HER) has been used to measure the effect of hydrogen flow control, which is defined in Equation (4).

An adequate and effective supply of hydrogen can be ensured by controlling the circulation flow of hydrogen to ensure the effective operation of the electrochemical reaction

while improving the utilization of hydrogen and economic efficiency [15]. The structure of the hydrogen circuit supply system is relatively simple, and its common control algorithms, including traditional control and predictive control methods, are able to control the inflow and outflow of the hydrogen circuit well. In Ref. [16], the fuel cell hydrogen gas pressure and HER were controlled by a classical proportional-integral differential controller and a state feedback controller, with good tracking and interference resistance results. In Ref. [17], a Markov chain and model predictive control (MPC) method was used to control the fuel cell anode supply system, which further improved the dynamic performance of the anode gas supply system response. Yao Wang [18] designed an adaptive backpropulsion sliding mode (ABSM) controller to control the output mass flow rate of a hydrogen circulation pump, and tested the effectiveness of the ABSM controller by establishing a nonlinear model of the hydrogen supply system and using vehicle driving cycle data from WVUSUB.

Based on the above discussion, this paper adopts a new control method to solve the air and hydrogen flow control problem of FCS. The proposed method has the advantages of enhanced dynamic response performance and stability performance, thereby improving the output performance and economic benefit of the system. The main contributions of this paper are as follows.

1. A simplified hybrid model environment of the fuel cell air and hydrogen circuits is built;
2. An optimal flow control strategy based on net power optimization is proposed;
3. Deep reinforcement learning controllers based on deterministic policy gradient are proposed to control the oxygen flow and hydrogen flow. The effect of decoupled and coupled controllers is compared;
4. A controller that integrates fuzzy PID and DRL algorithms is proposed, which has a faster dynamic response than traditional PID and more stable steady-state performance than DRL algorithms.

The structure of this paper is listed as follows: Section 2 presents the dynamic mathematical model of the fuel cell gas supply system. In Section 3, the deep reinforcement learning algorithm principle is presented. Section 4 describes the algorithm of controllers proposed in this paper. Simulation results and performance analysis are shown in Section 5. Finally, conclusions are given in Section 6.

2. Fuel Cell Gas Supply System

The PEMFC is an electrochemical device that directly converts the chemical energy of gas into electric energy and heat energy by electrochemical reaction. Hydrogen and oxygen (or air) are used as fuel and oxidant of the PEMFC, respectively. Under the action of the proton exchange membrane (PEM) and catalyst, the gases involved in the reaction undergo electrochemical reactions to generate electric energy and heat energy. A schematic diagram of the composition and working principle of a single PEMFC is shown in Figure 1.

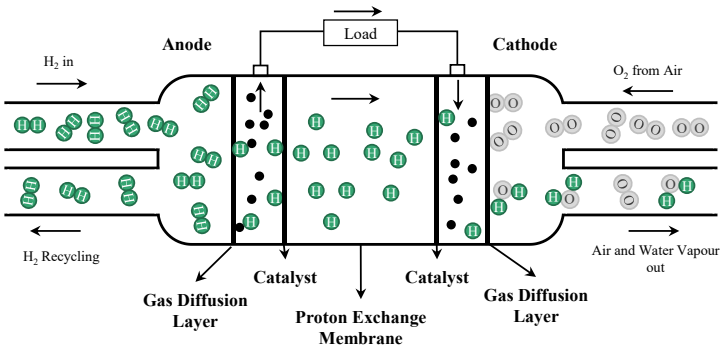

Figure 1. Electrochemical reaction of a fuel cell stack.

The PEM divides the PEMFC stack into anode and cathode. In the working process of the battery, the hydrogen at the anode loses two electrons under the action of the catalyst and is oxidized into H^+. The generated H^+ reaches the cathode through the PEM in the form of hydration, and the electrons reach the cathode after working on the load through the external circuit. After oxygen reaches the cathode, under the action of the cathode catalyst, it combines with H^+ and e^- reaching the cathode to generate water. Therefore, the total reaction process of the PEMFC can be described as:

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O$$

2.1. Fuel Cell Output Voltage Model

The output voltage model of the fuel cell reflects its electrical output performance. It is a function of the load current, the temperature in the stack, the partial pressure of the reactants, and the relative humidity of the gas. According to the thermodynamic equations in the standard state and the relevant thermodynamic data, the thermodynamic electromotive force of a single fuel cell in the ideal case can be expressed using the Nernst equation as [19]:

$$E = 1.229 - 0.85 \times 10^{-3}(T_{fc} - 298.15) + 4.3085 \times 10^{-5} T_{fc}\left[\ln(P_{H_2}) + \ln(P_{O_2})\right] \quad (1)$$

where P_{H_2} and P_{O_2} are the partial pressures of hydrogen and oxygen, respectively, in atm, and T_{fc} is the battery working temperature in K.

In the working process of the fuel cell, the polarization phenomenon will lead to some inevitable voltage loss, including activation polarization overvoltage V_{act}, ohmic overvoltage V_{ohm}, and concentration polarization overvoltage V_{conc}. The output voltage of a single fuel cell is shown in Equation (2), which is small. Therefore, multiple fuel cells are often connected in series to form a stack, and the overall output voltage of the fuel cell stack is obtained by Equation (3).

$$\begin{aligned} V_{fc} &= E - V_{act} - V_{ohm} - V_{conc} \\ &= E - \left(V_0 + V_a(1 - e^{-c_1 i})\right) - i\frac{t_m}{\sigma_m} - i(c_2\frac{i}{i_{\max}})^{c_3} \end{aligned} \quad (2)$$

where V_0 is the voltage at zero current density in V, i is the current density in A/m^2, i_{max} is the limit current density. i_{max}, c_1, c_2 and c_3 are the constants related to the gas pressure and temperature. t_m denotes the thickness of the PEM in m, and σ_m is the conductivity of the PEM, related to the temperature and membrane. The specific expressions of the relevant parameters are detailed in Ref. [20] and will not be repeated.

$$V_{stack} = nV_{fc} \quad (3)$$

where n represents the number of fuel cells in series.

2.2. Fuel Cell Gas Supply System Flow Mathematical Model

The PEMFC gas supply system includes hydrogen and air supplies. During the operation of the PEMFC, hydrogen and air enter the stack through the high-pressure hydrogen storage tank and air compressor, respectively, and flow along the gas flow channel in the stack to reach each cell. The reacting gas is electrochemically reacted by the electrode catalyst and the PEM, and the residual gas after the reaction is discharged from the stack through the regulating valve and other components. After a certain treatment, it can re-enter the stack for recycling. To better model the gas supply system, the following assumptions are made about the electrochemical reaction process of the fuel cell:

1. All gases are ideal gases;
2. Air flow, hydrogen flow, temperature, humidity, etc., are controlled separately;
3. The temperature inside the stack is uniformly distributed and always remains constant;

4. The gas pressure inside the stack is uniformly distributed;
5. The stack gas water vapor inside the stack is saturated; the liquid volume also has no effect on the system.

Based on the above assumptions and further neglecting the influence brought by the reaction gas transport pipeline, mathematical models and simulations of the dynamic description of the supply flow of air on the cathode side and hydrogen on the anode side can be obtained by the equation of state of ideal gas and the law of conservation of mass of matter, respectively.

HER and OER are used to measure the reasonableness of the anode hydrogen flow and the cathode oxygen flow, respectively, defined as the ratio of gas inflow and reaction:

$$\lambda_{H_2} = \frac{W_{H_2,in}}{W_{H_2,reacted}} \tag{4}$$

$$\lambda_{O_2} = \frac{W_{O_2,in}}{W_{O_2,reacted}} \tag{5}$$

where $W_{H_2,in}$ and $W_{H_2,reacted}$, $W_{O_2,in}$ and $W_{O_2,reacted}$ are the mass flows of hydrogen and oxygen into the stack and consumed by the electrochemical reaction in kg/s , respectively.

$$W_{H_2,reacted} = \frac{nI_{st}}{2F} M_{H_2} \tag{6}$$

$$W_{O_2,reacted} = \frac{nI_{st}}{4F} M_{O_2} \tag{7}$$

where M_{H_2} and M_{O_2} are the molar masses of hydrogen and oxygen in kg/mol, respectively. I_{st} is the stack current in A, F is Faraday constant with 96,487 C/mol.

2.3. Critical Component Models of the Fuel Cell Gas Supply System

In the process of controlling the parameters of the fuel supply effect, such as the oxygen ratio and hydrogen ratio of the fuel cell gas supply system, the elements that have a great influence on the control effect include the air compressor and circulating pump.

2.3.1. Air Compressor Model

The air compressor model is mainly based on the compressor rotation parameters and compressor air flow. The rotational speed dynamics of the air pressure can be modeled based on the rotational parameters:

$$J_{cp} \frac{d\omega_{cp}}{dt} = \tau_{cm} - \tau_{cp} \tag{8}$$

where J_{cp} is the rotational moment of inertia of the air compressor in kg·m^2, ω_{cp} is the rotational speed of the air compressor in rpm, and τ_{cm} and τ_{cp} are the motor driving torque and resistance torque of the air compressor in N·m, respectively.

The motor drive torque can be obtained by referring to the static equation of the motor:

$$\tau_{cm} = \eta_{cm} k_t i_{cm} \tag{9}$$

$$i_{cm} = \frac{v_{cm} - k_v \omega_{cp}}{R_{cm}} \tag{10}$$

where η_{cm} is the mechanical efficiency of the air compressor motor, k_t, k_v and R_{cm} are the drive motor constants, i_{cm} is the operating current in A, and v_{cm} is the motor operating voltage in V.

The resistance torque can be obtained by the thermodynamic equation:

$$\tau_{cp} = \frac{C_p T_{atm}}{\omega_{cp} \eta_{cp}} \left[\left(\frac{p_{sm}}{p_{atm}} \right)^{\frac{\gamma}{\gamma+1}} - 1 \right] W_{cp} \tag{11}$$

where C_p is the specific heat capacity constant of air in J/(kg·K), γ is the specific heat capacity at atmospheric pressure, W_{cp} is the air compressor air flow rate, and η_{cp} is the air compressor mechanical efficiency. P_{sm} and P_{atm} are the pressures of the supply pipe and outside air in kPa, respectively.

2.3.2. Recirculating Pump Model

The working principle of the recirculating pump is similar to that of the air compressor and will not be repeated. At the same time, its recirculating pump pressure ratio and power consumption are small, and the effect of power consumption is no longer considered in this paper. Therefore, this paper only considers the recirculating pump flow model, which is relatively simple and can be calculated by the difference in the data obtained offline, expressed as:

$$W_{rcp} = \frac{N_{rcp}}{N_0} W_0 \tag{12}$$

where W_{rcp} and W_0 are the flow rates at actual and rated speed in kg/s, respectively. N_{rcp} and N_0 are the actual and rated speed of the recirculating pump in rpm, respectively.

2.4. Fuel Cell Gas Supply System Model

Based on the available literature and the above introduction, a model of the fuel cell gas supply system can be built, as shown in Figure 2.

In this paper, the flow characteristics of air and hydrogen are mainly studied, and the main control elements are the air compressor, the recirculating pump, and some valves. The cooler and humidifier are used to maintain the temperature and humidity inside the cell, which are not studied. The relevant parameters come from Ref. [20], which are not repeated.

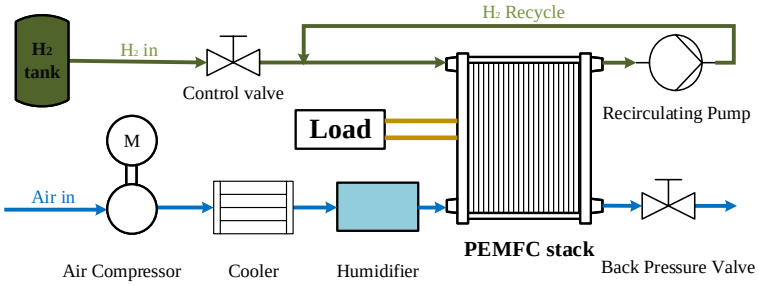

Figure 2. Fuel cell system structure diagram.

In addition, considering that in the gas supply system element, the power consumption of other part elements is much lower than that of the air compressor, the net power is simplified to Equation (13).

$$P_{net} = P_{st} - P_{cm} = I_{st} \cdot V_{stack} - i_{cm} \cdot v_{cm} \tag{13}$$

where P_{net}, P_{st} and P_{cm} represent the net power, output power, and air compressor power consumption of the system in W, respectively.

3. Deep Reinforcement Learning Algorithm

3.1. Deep Reinforcement Learning

David Silver has pointed out that 'deep learning + reinforcement learning = general artificial intelligence' [21]. Deep reinforcement learning (DRL) is the combination of deep learning (DL) and reinforcement learning (RL).

DL is a deep machine learning model. Its concept comes from an artificial neural network (ANN), and its depth is reflected in the multiple transformations of features. The commonly used deep learning model is a multilayer neural network. It can automatically learn abstract and essential features from a large amount of training data, realize the approximation of complex functions, and discover the distributed features of data [22,23].

RL is a special, mechanical learning method that uses model feedback as input and adapts to changes in the model (Environment). It is a method in which an intelligence (Agent) takes an action to interact with the environment to obtain a reward based on its states and finally completes an optimal policy to maximize the reward [24]. The basic RL model includes the strategy, reward function, value function, and environment model [25].

In DRL algorithms, DL refines the data through learning, while reinforcement learning is mainly used to solve the problem of time series decision-making. Through continuous interaction and trial and error with the environment, RL can finally obtain the optimal strategy of a specific task and maximize the cumulative expected return of the task.

According to the different ways that the deep neural network approaches nonlinear functions, DRL can be divided into value-based methods [26] and policy-based methods [27]. The value-based DRL uses deep neural network (DNN) to approximate the reward value function, and the most classical algorithm is the deep Q-learning (DQN) method, whose actions are discrete values, and the algorithm flow is shown in Algorithm 1. Approximating the policy with DNN and finding the optimal policy using the policy gradient method is known as policy-based DRL. The method of the policy gradient has advantages such as better convergence compared to the value function method, an easier strategy exploration process, and the ability to learn stochastic strategies.

Algorithm 1 DQN algorithm

Initialize action-value function Q random parameters θ_Q
Initialize target networks $\theta_{Q^*} \leftarrow \theta_Q$
Initialize replay buffer $\mathcal{B}$
for $episode = 1, ...M$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for $t = 1, ...T$ **do**
    Select action $a_t = \max_a Q^*(s_t, a \mid \theta_Q)$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{B}$
    Sample mini-batch of N transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{B}$
    Set $y_i = r_i + \gamma \max_{a'} Q(s_{i+1}, a^{'})$
    Update θ_Q by minimizing the loss: $\theta_Q \leftarrow \min_{\theta_Q} N^{-1} \sum_i (y_i - Q(s_i, a_i))^2$
  end
end

3.2. Principles of Deep Reinforcement Learning Algorithm Based on Deterministic Policy Gradient

Deep deterministic policy gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient (TD3), typical DRL algorithms based on deterministic policy gradients, are used as the control algorithm for fuel cell gas supply system in this paper.

In the training process, the DDPG algorithm involves four networks: actor current network π, actor target network π^*, critic current network Q, and critic target network Q^*. In addition, DDPG uses empirical playback for computing the target Q values.

TD3 is an extension algorithm of the DDPG algorithm, that proposes three key techniques based on DDPG:

1. Dual network: K (usually 2) sets of critic networks are used, and the smallest is taken when calculating the target value, thus suppressing the network overestimation problem;
2. Target policy smoothing regularization: when calculating the target value, a perturbation is added to the action of the next state, thus making the value evaluation more accurate;
3. Delayed update: the actor network is updated only after every d updates of the critic network, thus ensuring a more stable training of the actor network.

The overall algorithm flow for the training process of DDPG and TD3 is shown in Algorithm 2, where K and d are both 1 in DDPG and an integer greater than 1 in TD3.

Algorithm 2 DDPG/TD3 algorithm

Initialize K critic networks $Q_k(k = 1, 2..., K)$, and actor network π with random parameters θ_{Q_k}, θ_π
Initialize target networks $\theta_{Q_k^*} \leftarrow \theta_{Q_k}$, $\theta_{\pi^*} \leftarrow \theta_\pi$
Initialize replay buffer $\mathcal{B}$
for *episode* $= 1, ...M$ **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for $t = 1, ...T$ **do**
    Select action $a_t = \pi(s_t \mid \theta_\pi) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition tuple (s_t, a_t, r_t, s_{t+1}) in $\mathcal{B}$
    Sample mini-batch of N transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{B}$
    Set $y_i = r_i + \gamma Q^*(s_{i+1}, \pi^*(s_{i+1} \mid \theta_{\pi^*}) \mid \theta_{Q^*})$, where $Q^* = \min_{k=1,2,...K} Q_k$
    Update θ_{Q_k} by minimizing the loss: $\theta_{Q_k} \leftarrow \min_{\theta_{Q_k}} N^{-1} \sum_i (y_i - Q_k(s_i, a_i))^2$
    if t mod d **then**
      Update θ_π by the deterministic policy gradient:

$$\nabla_{\theta_\pi} \pi \mid_{s_i} = N^{-1} \sum_i \nabla_a Q_{\theta_1}(s_i, a) \mid_{a=\pi(s_i)} \nabla_{\theta_\pi} \pi(s_i)$$

      Update target networks:

$$\theta_{\pi^*} \leftarrow \tau\theta_\pi + (1 - \tau)\theta_{\pi^*}$$

$$\theta_{Q_k^*} \leftarrow \tau\theta_{Q_k} + (1 - \tau)\theta_{Q_k^*}$$

    end
  end
end

4. Deep Reinforcement Learning Algorithm Based on Fusion Optimization

4.1. Adaptive Expectations

In this paper, the variables related to the control objectives include the OER and the HER. In most studies, their optimal values are set to fixed values of 2 and 1.5 [28], respectively. However, in practice, their desired values are generally set to adaptive values related to the load demand to improve the hydrogen utilization and the output efficiency of the stack.

In this paper, the hydrogen system adopts a cyclic structure with a high hydrogen utilization rate [28]. Moreover, considering that the power consumption of the hydrogen

circuit components is much lower than that of the air compressor, it is no longer taken into account in the auxiliary system power consumption in order to simplify the calculation. Therefore, the expected value of the HER satisfies that the hydrogen demand and the pressure difference between the two poles of the stack can be maintained, so it is still set to 1.5, or $\lambda^*_{H_2} = 1.5$. For the hydrogen circulation system, the rotation speed of the circulation pump is used to control the circulation amount of hydrogen, and its expected value is shown in Equation (14), using Equations (6), (4) and (12).

$$N_{rcp,ref} = \left(\lambda^*_{H_2} - 1\right) \frac{(nI_{st}M_{H_2})/2F}{W_0} N_0 \tag{14}$$

For the oxygen supply system, the net power is mainly affected by the air mass flow rate and the system current during the actual operation, so to improve the efficiency and power output, the maximum net power is set as the target of the expected value of the OER. Figure 3a shows the relationship between the OER λ_{O_2} and net power at different current values from 100 A to 300 A by offline simulation experiments. At a certain load current, with the increase in λ_{O_2}, the net power of the system shows a trend of first increasing and then decreasing.

The control indicator function is obtained by nonlinearly fitting these optimal points (black points in Figure 3b), as shown in the red curve in Figure 3b. The final optimal control reference $\lambda^*_{O_2}$ is a function of the stack current I_{st}, shown as:

$$\lambda^*_{O_2} = 9.453 \times 10^{-7} I_{st}^2 - 0.002 I_{st} + 2.439 + 31.397 I_{st}^{-1} \tag{15}$$

Additionally, the control quantity related to the overshoot ratio is mainly the compressor voltage, and a stable value of the overshoot ratio under steady-state conditions shows a constant compressor input voltage. Therefore, the value of the compressor voltage corresponding to the best OER $\lambda^*_{O_2}$ in Equation (15) for different load current demands is obtained by offline training, as shown in Figure 3c.

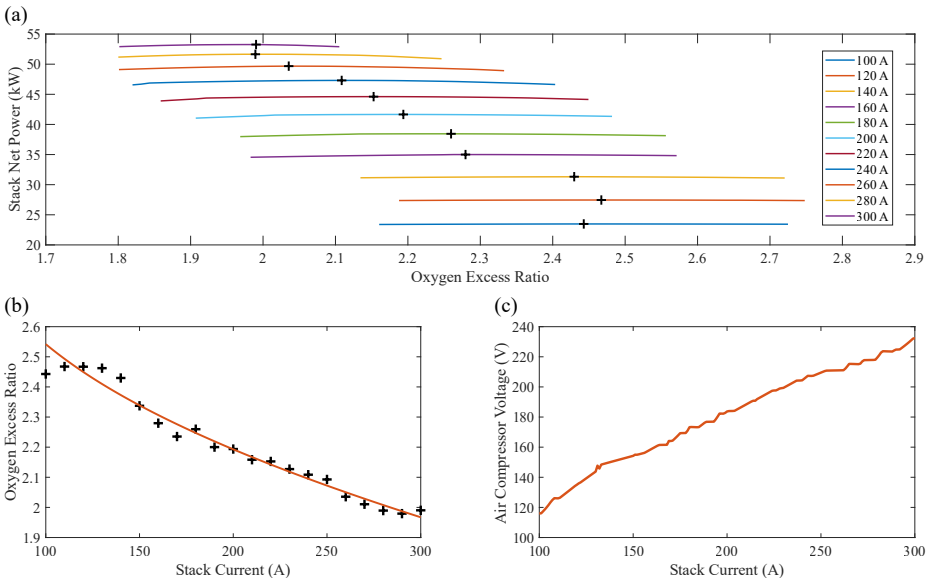

Figure 3. Optimum OER and air compressor voltage under different stack currents, where the black points indicate the optimal points: (**a**) the relationship between OER and the net power output of the fuel cell system under different current requirements, (**b**) the optimal OER under different current, (**c**) the optimal air compressor voltage under different current.

4.2. Deep Reinforcement Learning Controller

The integrated controller for the gas supply system described in this paper has three main control objectives: (1) to ensure proper OER or air flow; (2) to ensure proper HER or hydrogen flow; (3) to ensure the stability of the pressure difference between the two poles of the stack. To further simplify the complexity of the control algorithm, a follower control algorithm is used for the control of the two-pole pressure difference, resulting in a certain coupling between the control of the hydrogen circuit and the oxygen circuit.

Therefore, the structures of the DDPG and TD3 controller used in this paper are shown in Figure 4, whose agent consists of an actor and a critic network. The DQN controller is similar to Figure 4, except that the agent does not contain the critic network, which is not drawn here. The controller obtains the optimal OER and HER according to the demand load current typed by the outside world and takes the fuel cell gas supply system as the environment of the controller.

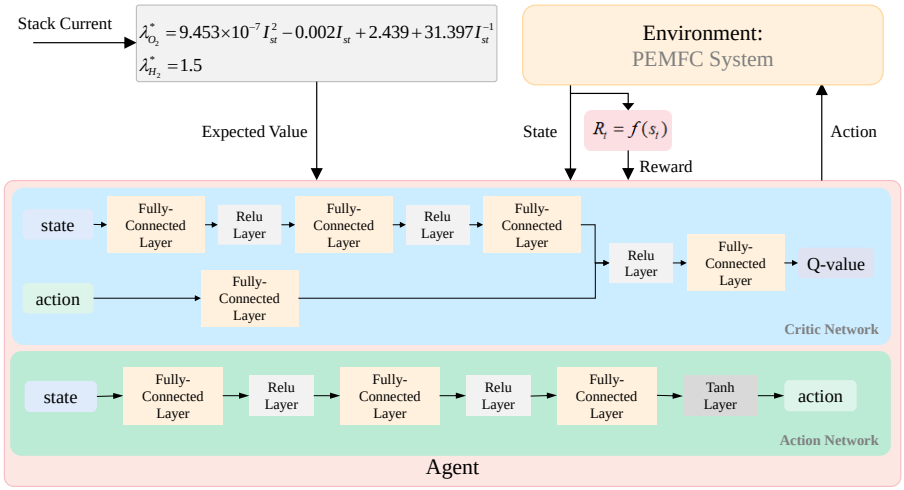

Figure 4. Structure of the DRL algorithms.

To better control the two-way gas supply, decoupled and uncoupled DRL control algorithms are designed, respectively. In the decoupled controller, called 2-DRL, there are two agents, which are used to control the hydrogen and the air circuit, respectively. However, in the uncoupled DRL controller, only one agent is used to control two gas supplies at the same time. The relevant parameter settings of the two algorithms are shown in Table 1 and the rewards are defined in Equations (16)–(18). Besides, to ensure that the overall complexity of the two DRLs is similar, the network structure of the two agents in 2-DRL is consistent with that in DRL and the number of neurons in each layer is half of that in DRL.

Table 1. Action, state, and reward settings in DRL algorithms.

Parameters	**2-DRL**		**DRL**
	agent_O_2	**agent_H_2**	
action	$V_{cm}(a_1)$	$N_{rcp}(a_2)$	V_{cm}, N_{rcp}
state	I_{st}, e_{OER}, Δe_{OER}, V_{cm}, ΔV_{cm}	I_{st}, e_{HER}, Δe_{HER}, N_{rcp}, ΔN_{rcp}	I_{st}, e_{OER}, Δe_{OER}, V_{cm}, ΔV_{cm}, e_{HER}, Δe_{HER}, N_{rcp}, ΔN_{rcp}
reward	$R_{OER} + R_{a_1}$	$R_{HER} + R_{a_2}$	$R_{OER} + R_{a_1} + R_{HER} + R_{a_2}$

The setting of the reward function requires several factors to be considered. The first consideration is the effect of the controlled amounts, λ_{H_2} and λ_{O_2}. To ensure that the error

of the controlled quantity gradually decreases and converges to 0, errors (e_{OER}, e_{HER}) and their variation($\Delta|e_{\mathrm{OER}}|$, $\Delta|e_{\mathrm{HER}}|$) need to be considered, as shown in Equations (16) and (17).

$$R_{\mathrm{OER}} = \begin{cases} 100(0.1 - |e_{\mathrm{OER}}|) + 9, & |e_{\mathrm{OER}}| < 0.1 \\ 10(1 - |e_{\mathrm{OER}}|), & 0.1 \le |e_{\mathrm{OER}}| < 1 \\ -1(|e_{\mathrm{OER}}| - 1), & 1 \le |e_{\mathrm{OER}}| < 2 \\ -0.1(|e_{\mathrm{OER}}| - 2) - 1, & \text{otherwise} \end{cases} + \begin{cases} 1, & \Delta|e_{\mathrm{OER}}| < \delta \\ -0.2\Delta|e_{\mathrm{OER}}|, & \text{otherwise} \end{cases} \tag{16}$$

$$R_{\mathrm{HER}} = \begin{cases} 100(0.1 - |e_{\mathrm{HER}}|) + 8, & |e_{\mathrm{HER}}| < 0.1 \\ 20(0.5 - |e_{\mathrm{HER}}|), & 0.1 \le |e_{\mathrm{HER}}| < 0.5 \\ -1(|e_{\mathrm{HER}}| - 0.5), & 0.5 \le |e_{\mathrm{HER}}| < 1.5 \\ -0.1(|e_{\mathrm{HER}}| - 1.5)) - 1, & \text{otherwise} \end{cases} + \begin{cases} 1, & \Delta|e_{\mathrm{HER}}| < \delta \\ -0.2\Delta|e_{\mathrm{HER}}|, & \text{otherwise} \end{cases} \tag{17}$$

where δ represents a very small positive number.

In addition, to avoid the problem of the action amount changing too much in the actual working process, the influence of the action amount changing value is also considered in the reward function.

$$R_{a_i} = \begin{cases} 1, & |\Delta a_i| < 0.01 \\ -0.1|\Delta a_i|, & \text{otherwise} \end{cases} \tag{18}$$

where $a_i (i = 1, 2)$ represents the normalized action values of V_{cm} and N_{rcp} respectively.

4.3. Deep Reinforcement Learning Controller Based on Fusion Optimization

To improve the steady-state performance of the controller, a method of DRL and fuzzy PID fusion is introduced. The overall structure is shown in the Figure 5. Among them, the input of the fuzzy logic in the fuzzy PID controller is the error and error variation of OER and HER, and the output is the PID control parameters, as shown in Figure 6. The fusion algorithm adopts the strategy of fuzzy logic, whose input is as shown in Figure 6 and the output is the weight of the fuzzy PID, as shown in Figure 7.

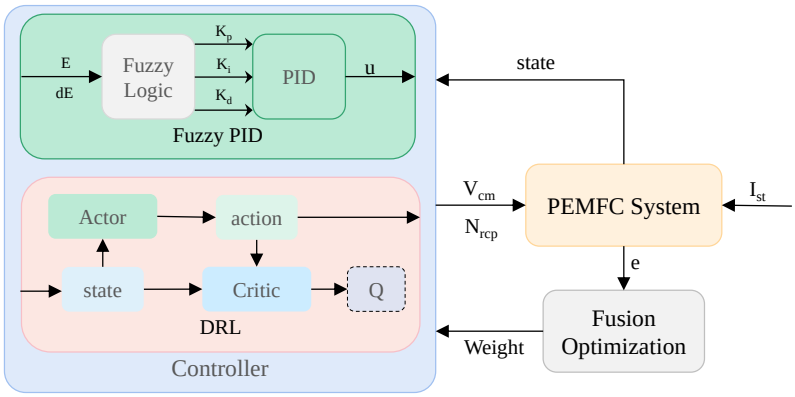

Figure 5. Overall structure of the controller.

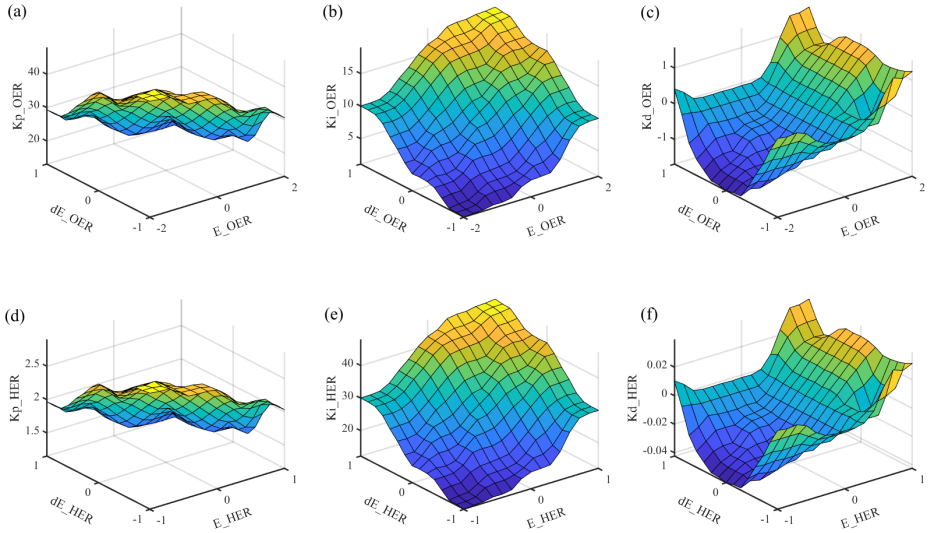

Figure 6. Fuzzy PID parameters: (**a**–**c**), (**d**–**f**) represent the PID parameters Kp, Ki, and Kd of the fuzzy logic output of OER and HER, respectively.

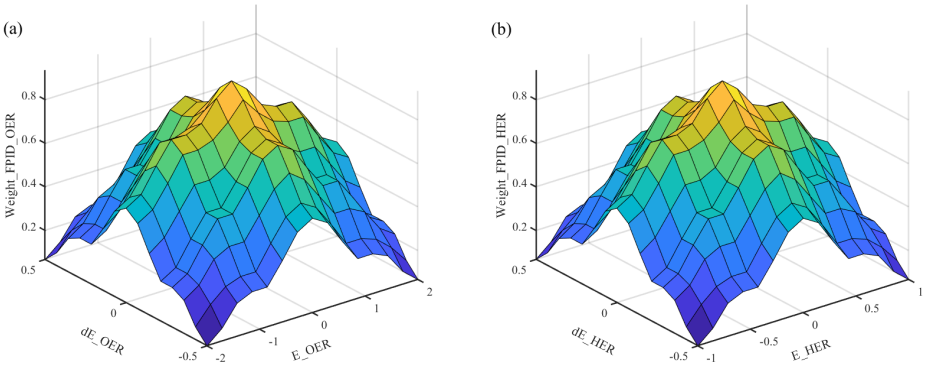

Figure 7. The weight value of FPID output by fuzzy logic: (**a**) OER control, (**b**) HER control.

5. Simulation Results and Analysis

5.1. DRL Controller Training and Testing Results

For the training of the DRL controllers of the fuel cell gas supply system, the relevant parameters used are shown in Table 2. In the training process, to improve the adaptability of the controller to different load currents, the input to the training is a step load current of random amplitude (100–300 A), and the training time is 10 s per cycle.

Table 2. DRL training parameters.

Symbol	Value	Unit	Instructions
M	1000	-	Maximum number of cycles
T_s	0.01	s	Sampling period
T	1000	s	Maximum steps per cycle
γ	0.9	-	Discount factor
τ	5×10^{-4}	-	Learning rate

To better compare the control effects of the DQN, DDPG, and TD3 algorithms on the two gas supplies, the DRL algorithm is used to control the air and hydrogen loops separately, and the changes in the average reward values during training are shown in Figure 8. The results show that DDPG has the relatively best control effect for both gases,

and its convergence values are close to the expected value (21,000/20,000). DQN has a better control effect for both gases, although it can reach a certain large expected value, but not as good as DDPG, and the convergence values of hydrogen circuit have a larger gap. This is due to the discrete action of action in DQN and the possible oscillation of the control effect under static conditions. TD3 has a better control effect for the air path and requires a shorter number of cycles to reach convergence, while in the hydrogen circuit, the training stability value of TD3 is much smaller than its expected value, which is because the TD3 network has two critic networks. Although too many evaluation networks improve the generalization ability of the controller, the accuracy of the simpler model is also reduced, leading to unsatisfactory control of the hydrogen circuit. Therefore, TD3 correlation will not be used as the control algorithm for the hydrogen circuit in the subsequent study.

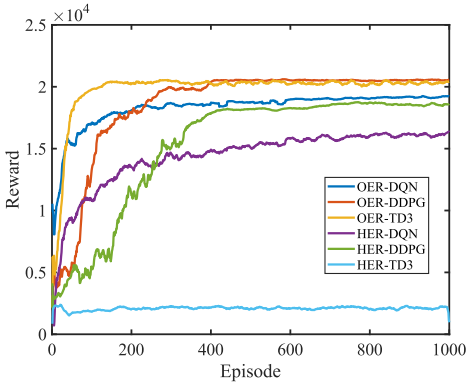

Figure 8. Reward values during training of OER and HER individual DRL controllers.

Therefore, to better compare the control effects of decoupled and uncoupled, DQN, DDPG, and TD3 networks, the average reward value variation during training is further designed to compare the uncoupled DQN and DDPG algorithms, as well as the decoupled 2DQN, 2DDPG, DDPG-DQN, (where DQN is used for oxygen and DDPG for hydrogen circuit) and DDPG-TD3 (where TD3 is used for oxygen and DDPG for hydrogen circuit), as shown in Figure 9. The overall reward value in the decoupled DRL controller is the sum of the reward values in the two agents.

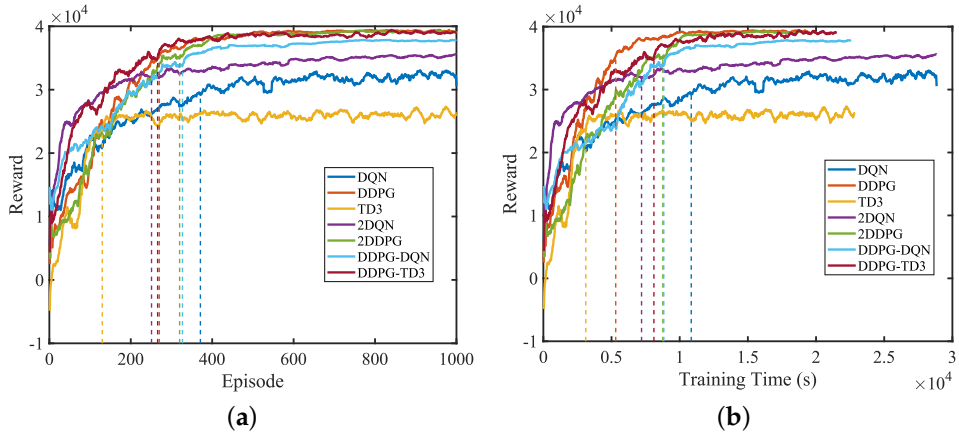

Figure 9. Reward values during DRL controller training. (**a**) Reward changes in relation to episode. (**b**) Reward changes in relation to time.

As seen, the DDPG, 2DDPG, DDPG-DQN, and DDPG-TD3 algorithms can all converge the controller's reverse to the desired value (41,000), and the training cycles they need to reach 90% of the desired value do not differ much, but DDPG has a faster training speed and takes less time to reach the steady state. 2TD3 and TD3 algorithms have a shorter training

rise time than other algorithms, but their training final steady-state values are smaller, which indicates that the hydrogen supply training is still very poor and cannot be used as a control algorithm. For the DQN and 2DQN, the convergence stability values have a large gap with the expected value, which indicates that the overall control effect of the controllers cannot achieve the expected effect well. This is mainly because the action of DQN is discrete and its action value may oscillate around the optimal value in the steady-state operating condition. Additionally, considering that the 2DQN and DDPG-DQN, DDPG and TD3 are similar in structure, only the DQN, DDPG, MDQN (DDPG-DQN), and MTD3 (DDPG-TD3) controllers are studied for detailed comparison in the subsequent study.

5.2. Controller Comparison Test and Result Analysis

After controller training, the control effect is verified using random load currents, and the set current values are shown in Figure 10.

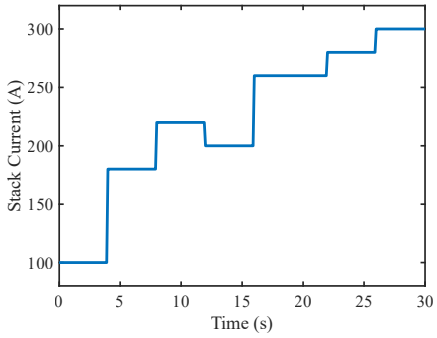

Figure 10. Test stack current profile.

To further demonstrate the control performance of the proposed method, the control results of the two other control methods are discussed and compared. The first is feed-forward open-loop control, whose prior control quantity is determined by the stack current, calculated by Figure 3c and Equation (14). The second method is fuzzy PID control, whose parameter settings are as shown in Figure 6. Figures 11 and 12 show the control results and expected control values for the control quantities and in each of the six methods.

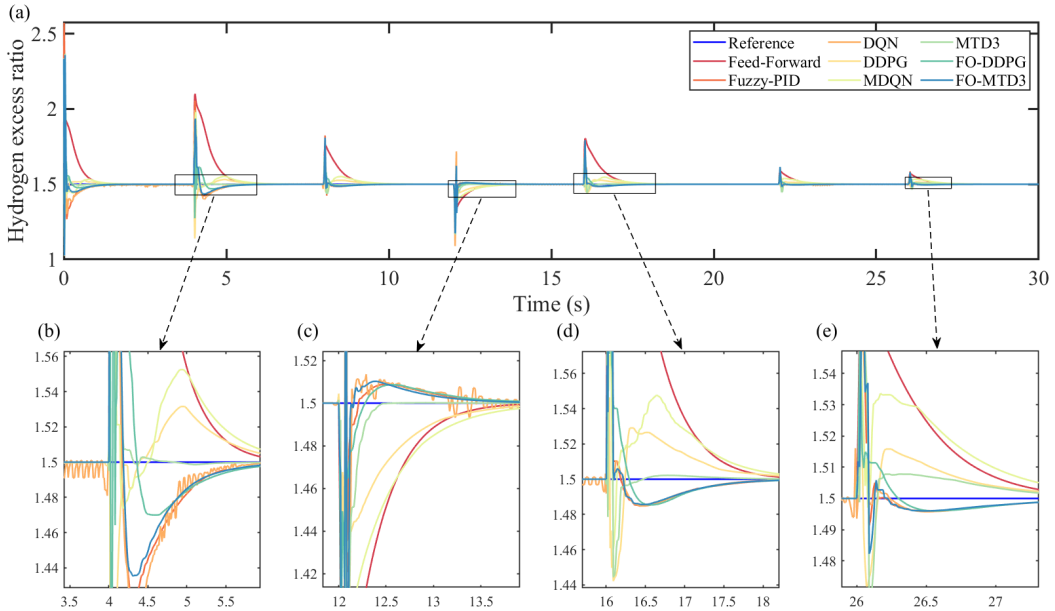

Figure 11. Test results of hydrogen excess ratio control: (**b**–**e**) are the enlarged figures of (**a**) near 4 s, 12 s, 16 s, 26 s when the load current changes, respectively.

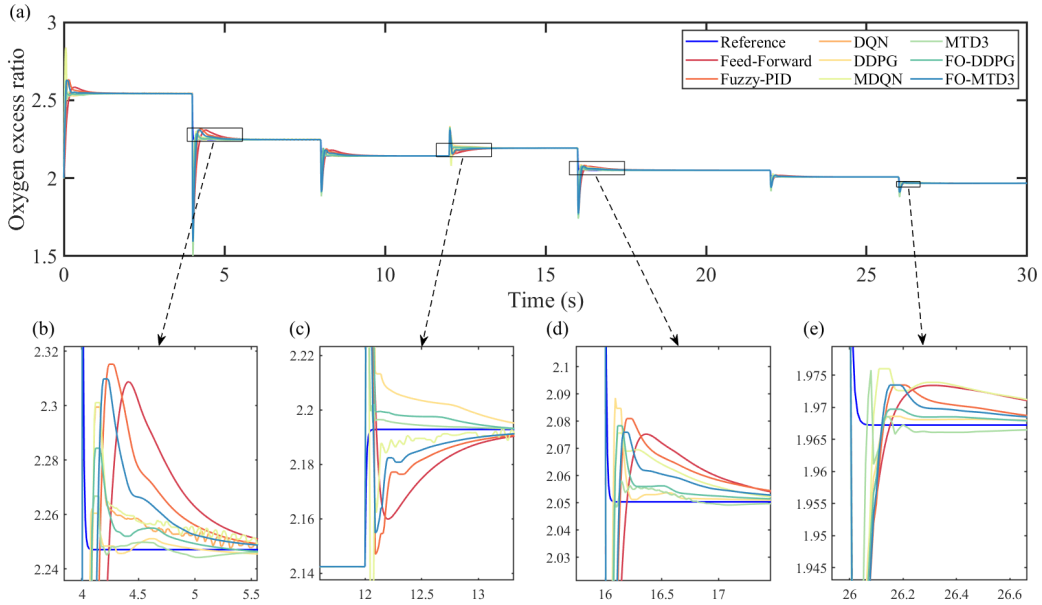

Figure 12. Test results of oxygen excess ratio control: (**b**–**e**) are the enlarged figures of (**a**) near 4 s, 12 s, 16 s, 26 s when the load current changes, respectively.

In addition, four parameters are used to measure the control effect of the controllers: the rise time T_r (error less than 5%), the maximum overshoot M_d, the root mean square error within 2 s RMSE_{+2s} after the abrupt change of operating conditions, and the root mean square error RMSE_{st} when the system reaches stability, which are defined in Equations (19) and (20). The calculated results are shown in Table 3, and the value marked in red represents the optimal effect value of a control algorithm in a variety of control algorithms.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i,\mathrm{real}} - x_{i,\mathrm{ref}}\right)^2} \tag{19}$$

$$M_d = \max_i\left(\left|\frac{x_{i,\mathrm{real}} - x_{i,\mathrm{ref}}}{x_{i,\mathrm{ref}}}\right|\right) \times 100\% \tag{20}$$

where n is the number of sampling steps of the whole process, and $x_{i,\mathrm{real}}$ and $x_{i,\mathrm{ref}}$ are the actual and expected values of the ith sampling time, respectively.

Table 3. Control effect parameter values of multiple algorithms, where the red values indicate the best values for each algorithm in the table.

Algorithm	HER				OER			
	RMSE_{+2s}	RMSE_{st}	T_r(s)	M_d(%)	RMSE_{+2s}	RMSE_{st}	T_r(s)	M_d(%)
Feed-Forward	0.1215	0.000046	0.89	31.22%	0.0611	0.000052	0.16	2.90%
Fuzzy-PID	0.0699	0.000033	0.36	9.09%	0.0561	0.000014	0.12	3.20%
DQN	0.0643	0.011157	0.28	8.28%	0.0485	0.004120	0.11	1.02%
DDPG	0.0386	0.005113	0.15	2.44%	0.0420	0.005646	0.11	0.64%
MDQN	0.0508	0.005309	0.26	5.64%	0.0481	0.004019	0.11	1.04%
MTD3	0.0519	0.003132	0.71	5.95%	0.0472	0.000828	0.07	0.40%
FO-DDPG	0.0370	0.000033	0.15	4.36%	0.0441	0.000031	0.11	1.01%
FO-MTD3	0.0545	0.000035	0.24	7.30%	0.0487	0.000010	0.10	0.92%

The experimental results show that DQN and MDQN are far inferior to DDPG and MTD3. The MTD3 algorithm has good dynamic characteristics, short rise time, and small

overshoot for oxygen supply control, while the coupled DDPG algorithm has good dynamic control for hydrogen supply; when considering the hysteresis of the air compressor, in the control of two-stage differential pressure, it is the anode that follows the cathode, i.e., the hydrogen inflow. HER controlled by the regulating valve on the hydrogen circuit will be affected by the state of the system on the air circuit. In the decoupled control of gas on both sides, the state of the other side is generally not considered, so HER works better in the coupled control. The air circuit belongs to the autonomous control, which is less influenced by the cathode, that is, the state of hydrogen circuit is an irrelevant variable for OER control, so OER is better controlled in the decoupling control, but still not too much better than the coupling control. Although the common DRL algorithm can effectively improve the dynamic response speed of the system, its steady-state performance is far inferior to the common feedforward and fuzzy PID control. FO-DDPG integrates the advantages of DDPG and PID algorithms, improves the dynamic response speed and steady-state performance of the system, and makes the system have better robustness and stability. In a comprehensive comparison, the FO-DDPG algorithm has an overall better control performance.

6. Conclusions

In this paper, a DRL controller with continuous state based on fusion optimization was proposed for a PEMFC gas supply system. By controlling the in and out of air and hydrogen, the combined control of two gas flows in the fuel cell gas supply system was realized. A control strategy based on net power optimization and high economy was also introduced in the experiments, and the need for decoupled control of the two gases was also investigated. These control algorithms were compared with feedforward controllers, fuzzy PID and discrete DRL control algorithms DQN, and conventional continuous DRL algorithms DDPG.

The experimental results show that the FO-DDPG algorithm proposed in this paper has a faster dynamic response and stable static performance compared to the traditional fuzzy PID, DQN, and DDPG algorithms. Specifically, for the two controlled quantities, the dynamic response time is only 0.1 5 s, the overshoot is less than 5%, and the stability error of the control quantity is only about 0.00003. That is, the controller can meet the real-time control requirements of the gas supply system under different load conditions, effectively avoiding the undersupply or oversupply of oxygen and hydrogen, and improving the power and economic efficiency of the system.

PEMFC is a highly complex system involving many components, and many areas need to be controlled. The research in this paper is limited to ensuring sufficient supply of the reaction gas and is somewhat ideal for environmental settings. In a real fuel cell system, factors such as temperature and humidity inside the stack can affect the electrochemical reaction and the output voltage of the fuel cell, and the bipolar pressure can also affect the transport of the PEM. In future research, other meta-components can be introduced to further improve the efficiency of the fuel cell by constructing a multi-physical quantity multi-dimensional fuel cell model and designing a gas-water-heat-electric coordinated integrated controller to realize the joint control of power, pressure, temperature, and humidity of the fuel cell.

Author Contributions: Conceptualization, H.Y. and Z.C.; Methodology, H.Y., Z.S. and Z.C.; Software, H.Y. and Z.S.; Formal analysis, Z.C.; Resources, Z.C.; Writing—original draft, H.Y.; Writing—review & editing, Z.S., Y.W. and Z.C.; Supervision, Z.C.; Funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Kojima, K. Recent progress of research and development for fuel cell vehicle application. In Proceedings of the 2008 International Conference on Effects of Hydrogen on Materials, Grand Teton National Park, WY, USA, 7–10 September 2008.
2. Wang, Y.; Chen, K.S.; Mishler, J.; Cho, S.C.; Adroher, X.C. A review of polymer electrolyte membrane fuel cells: Technology, applications, and needs on fundamental research. *Appl. Energy* **2011**, *88*, 981. [CrossRef]
3. Liu, J.; Zhou, Z.; Zhao, X.; Xin, Q.; Sun, G.; Yi, B. Fuel Cell Overview. *Phys. Chem. Chem. Phys.* **2004**, *6*, 134. [CrossRef]
4. Lachaize, J.; Caux, S.; Fadel, M.; Schott, P.; Nicod, L. Pressure, flow and thermal control of a fuel cell system for electrical rail transport. In Proceedings of the 2004 IEEE International Symposium on Industrial Electronics, Ajaccio, France, 4–7 May 2004.
5. Goshtasbi, A.; Ersal, T. Degradation-conscious control for enhanced lifetime of automotive polymer electrolyte membrane fuel cells. *Power Sources* **2020**, *457*, 227996. [CrossRef]
6. Ryu, S.K.; Vinothkannan, M.; Kim, A.R.; Yoo, D.J. Effect of type and stoichiometry of fuels on performance of polybenzimidazole-based proton exchange membrane fuel cells operating at the temperature range of 120–160 C. *Energy* **2022**, *238*, 121791. [CrossRef]
7. Matraji, I.; Laghrouche, S.; Wack, M. Pressure control in a PEM fuel cell via second order sliding mode. *Int. J. Hydrogen Energy* **2012**, *37*, 16104. [CrossRef]
8. Matraji, I.; Laghrouche, S.; Wack, M. Second order sliding mode control for PEM fuel cells. In Proceedings of the 49th IEEE Conference on Decision and Control, Atlanta, GA, USA, 15–17 December 2010.
9. Tang, X.; Wang, C.S.; Mao, J.H.; Liu, Z.J. Adaptive fuzzy PID for proton exchange membrane fuel cell oxygen excess ratio control. In Proceedings of the 32nd Chinese Control and Decision Conference, Hefei, China, 22–24 August 2020.
10. Wei, G.; Quan, S.; Zhu, Z.; Pan, M.; Qi, C. Neural-PID control of air pressure in fuel cells. In Proceedings of the 2010 International Conference on Measuring Technology and Mechanical Automation(ICMTMA), Changsha, China, 13–14 March 2010.
11. Kim, E.S.; Kim, C.J. Nonlinear State Space Model and Control Strategy for PEMFC systems. *J. Energy Power Eng.* **2010**, *4*, 8.
12. Yang, D.; Pan, R.; Wang, Y.; Chen, Z. Modeling and control of PEMFC air supply system based on TS fuzzy theory and predictive control. *Energy* **2019**, *188*, 116078. [CrossRef]
13. Sedighizadeh, M.; Rezazadeh, A. Adaptive Self-Tuning Wavelet Neural Network Controller for a Proton Exchange Membrane Fuel Cell. *Appl. Neural Netw. High Assur. Syst.* **2010**, *268*, 221–245.
14. Li, C.; Zhu, X.; Sui, S.; Hu, W.; Hu, M. Maximum power point tracking of a photovoltaic energy system using neural fuzzy techniques. *J. Shanghai Univ.* **2009**, *13*, 29–36. [CrossRef]
15. Hwang, J.J. Effect of hydrogen delivery schemes on fuel cell efficiency. *J. Power Sources* **2013**, *239*, 54. [CrossRef]
16. He, J.L.; Choe, S.Y.; Hong, C.O. Analysis and control of a hybrid fuel delivery system for a polymer electrolyte membrane fuel cell. *J. Power Sources* **2008**, *185*, 973. [CrossRef]
17. Quan, S.W.; Chen, J.Z.; Wang, Y.X.; He, H.; Li, J. A hierarchical predictive strategy-based hydrogen stoichiometry control for automotive fuel cell power system. In Proceedings of the 16th IEEE Vehicle Power and Propulsion Conference, Hanoi, Vietnam, 14–17 October 2019.
18. Wang, Y.; Quan, S.; Wang, Y.; He, H. Design of adaptive backstepping sliding mode-based proton exchange membrane fuel cell hydrogen circulation pump controller. In Proceedings of the Asia Energy and Electrical Engineering Symposium, Chengdu, China, 28–31 May 2020.
19. Lee, J.H.; Lalkb, T.R.; Appleyc, A.J. Modeling electrochemical performance in large scale proton exchange membrane fuel cell stacks. *J. Power Sources* **1998**, *70*, 2. [CrossRef]
20. Pukrushpan, J.T.; Stefanopoulou, A.G.; Peng, H. *Control of Fuel Cell Power Systems*; Springer: London, UK, 2004.
21. Silver, D. Deep Reinforcement Learning. A Tutorial at ICML 2016, 19 June 2016. Available online: https://www.deepmind.com/learning-resources/introduction-to-reinforcement-learning-with-david-silver (accessed on 16 January 2023).
22. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504. [CrossRef] [PubMed]
23. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
24. Gao, Y.; Chen, S.F.; Lu, X. Research on reinforcement learning technology: A review. *Acta Autom. Sin.* **2004**, *30*, 86.
25. Montague, P.R. Reinforcement learning: An introduction. *Trends Cogn. Sci.* **1999**, *3*, 360. [CrossRef]
26. Boyan, J.A. Least-squares temporal difference learning. In Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 27 June 1999.
27. Aissani, N.; Beldjilali, B.; Trentesaux, D. Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach. *Eng. Appl. Artif. Intell.* **2009**, *22*, 1089. [CrossRef]
28. Rodatz, P.; Tsukada, A.; Mladek, M.; Guzzella, L. Efficiency improvements by pulsed hydrogen supply in PEM fuel cell systems. In Proceeding of the 15th IFAC Triennial World Congress, Barcelona, Spain, 21–26 July 2002.