*mathematics*

MDPI

*Article*

# A Novel Two-Stage, Dual-Layer Distributed Optimization Operational Approach for Microgrids with Electric Vehicles

**Bowen Zhou [1,2,*], Zhibo Zhang [1,2,*], Chao Xi [3] and Boyu Liu [4]**

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[2] Key Laboratory of Integrated Energy Optimization and Secure Operation of Liaoning Province, Northeastern University, Shenyang 110819, China
[3] State Grid Harbin Power Supply Company, Harbin 150001, China; 2070961@stu.neu.edu.cn
[4] School of Electrical Engineering and Telecommunications, UNSW Sydney, Sydney, NSW 2052, Australia; boyu.liu@unsw.edu.au
* Correspondence: zhoubowen@ise.neu.edu.cn (B.Z.); 2100690@stu.neu.edu.cn (Z.Z.)

**Abstract:** As the ownership of electric vehicles (EVs) continues to rise, EVs are becoming an integral part of urban microgrids. Incorporating the charging and discharging processes of EVs into the microgrid's optimization scheduling process can serve to load leveling, reducing the reliance of the microgrid on external power networks. This paper proposes a novel two-stage, dual-layer distributed optimization operational approach for microgrids with EVs. The lower layer is a distributed control layer, which ensures, through consensus control methods, that every EV maintains a consistent charging/discharging and state of charge (SOC). The upper layer is the optimization scheduling layer, determining the optimal operational strategy of the microgrid using the multiagent reinforcement learning method and providing control reference signals for the lower layer. Additionally, this paper categorizes the charging process of EVs into two stages based on their SOC: the constrained scheduling stage and the free scheduling stage. By employing distinct control methods during these two stages, we ensure that EVs can participate in the microgrid scheduling while fully respecting the charging interests of the EV owners.

**Keywords:** microgrid; electric vehicles; consensus control; deep reinforcement learning; microgrid optimization scheduling

**MSC:** 68U01; 68U35



**Citation:** Zhou, B.; Zhang, Z.; Xi, C.; Liu, B. A Novel Two-Stage, Dual-Layer Distributed Optimization Operational Approach for Microgrids with Electric Vehicles. *Mathematics* **2023**, *11*, 4563. https://doi.org/10.3390/math11214563

Academic Editor: Nicu Bizon

Received: 4 October 2023
Revised: 2 November 2023
Accepted: 3 November 2023
Published: 6 November 2023



## 1. Introduction

EVs, as clean and efficient means of transportation, not only enhance residents' mobility efficiency but also curtail urban pollutant emissions. Consequently, they have garnered extensive public acclaim. With the progressive refinement of EV technology in recent years, the ownership of EVs in urban areas has surged, making these vehicles an integral component of urban microgrids [1]. By judiciously scheduling EVs' charging and discharging processes, urban clean energy can be absorbed, thus decreasing the city's reliance on traditional energy sources. In addition, this simultaneously cuts the operational costs of urban electric systems, achieving cleaner and more cost-effective energy utilization [2]. Nowadays, as the call for sustainable urban development amplifies [3,4], EVs, as a novel form of energy provision, have provided fresh insights for energy transitions in cities globally, thus receiving escalating attention and support [5].

For instance, the European Union (EU) aspires to attain net-zero emissions by 2035. It has set forth plans to promote EVs by offering policy and financial incentives and building charging infrastructure across member states [6,7]. In 2021, the Indian government announced a hike in the subsidy for electric two-wheelers from 10,000 INR/kWh to 15,000 INR/kWh. It permitted EV manufacturers to offer up to a 40% discount to consumers [8].

The U.S. government in 2021 introduced incentives for EV deployment in 34 states, including exemptions for high-occupancy vehicle (HOV) lanes, financial perks for purchasing EVs or EV supply equipment, exemptions from vehicle inspections or emissions tests, parking incentives, and reduced electric rates for off-peak EV charging, among others [9].

Optimal charging/discharging scheduling research for EVs has emerged as a pivotal direction in the evolution of EV technology. Through optimizing the charging and discharging processes, stress on the electrical grid can be alleviated, facilitating sustainable energy utilization. In the foreseeable future, such optimal scheduling technologies will play an increasingly crucial role in the promotion and widespread adoption of EVs. Evolutionary algorithms and swarm intelligence optimization algorithms are considered effective methods for optimizing the charging and discharging process of EVs. Ref. [10] proposes a multiobjective optimization operation method based on the nondominated sorting genetic algorithm II (NSGA-II) for microgrids containing EVs. This method utilizes the charging and discharging of EVs to reduce the peak-to-valley value of $CO_2$ emissions and bus power in microgrids while considering the income of EV users. Ref. [11] introduces a scheduling model for EVs' charging based on swarm intelligence algorithms. This model contemplates the dynamic nature of charging demands and the uncertainty of user preferences, ultimately achieving a balance in the power system and maximizing user satisfaction. Ref. [12] proposes a two-stage ordered charging and discharging strategy for EVs using the particle swarm optimization (PSO) algorithm. Experimental results indicate that this approach enables more efficient utilization of grid resources during the EV charging and discharging stages, curtails peak charging loads, and enhances energy utilization and stability of the grid. Although evolutionary algorithms and swarm intelligence optimization algorithms can calculate the optimal decision-making process for charging and discharging EVs, they can only calculate a fixed decision-making process based on predetermined environmental conditions. For actual microgrids containing EVs, environmental conditions and EV charging session types are often uncertain [13]. Therefore, when evolutionary algorithms and swarm intelligence optimization algorithms are applied to real-world microgrids containing EVs, they often fail to achieve the expected results. To improve the robustness of microgrid operation strategies in practical environments, the structure of hierarchical learning, optimization, and control has been widely proposed. Ref. [14] and Ref. [15] proposed a hierarchical fuzzy control method that utilizes the idea of fuzzy control to ensure the robustness of the method. At the same time, by using a hierarchical architecture, the method ensures high accuracy while ensuring computational complexity. Ref. [16] proposes a multilevel game-theoretic model, where the upper level seeks to minimize the networkwide energy costs, and the lower level determines the optimal charging and discharging strategy for each EV by balancing cost minimization and revenue maximization. Experimental findings suggest that when considering demand response mechanisms, this model can elevate the energy efficiency of the power grid and augment the economic benefits for both EVs and the grid. Ref. [17] presents an EV charging coordination optimization method rooted in hierarchical optimization and user satisfaction. Considering the intricacies of the medium/low-voltage integrated network and factors like charging needs and user preferences, this method has proven to heighten user satisfaction while diminishing the strain on the power system. In Ref. [18], a rapid optimization algorithm is proposed to address multiple EVs' combined routing and charging problems. Experiments underscore that this algorithm can tackle numerous EVs' combined routing and charging problems, showing commendable time efficiency and solution quality. Ref. [19] proposes a hybrid integer linear programming model based on a virtual pricing mechanism to optimize grid energy efficiency while minimizing EV charging and discharging costs alongside user travel expenses. Studies illustrate that this model, while ensuring the travel needs of EVs, can actualize optimal energy allocation and load balance. Ref. [20] proposes an intelligent charging and discharging strategy for EVs in smart grids rooted in a decision function. This strategy can dynamically adjust EV charging and discharging timings and power according to electricity prices, grid load, and the charging needs of EV owners, maximizing

both owner benefits and grid advantages. Lastly, [21] proposes a real-time method to control EV charging and discharging in response to variations in renewable energy production and EV battery states. Research indicates that this technique can effectively govern EV charging and discharging, curtail energy consumption, and enhance the operational efficiency and reliability of charging stations and the grid.

While the aforementioned studies offer reasonable approaches to the optimized operation of EVs, there are several salient issues that cannot be overlooked:

1. Scheduling strategies are uniquely designated to each EV, making each vehicle's charging and discharging status independent of others. This approach can result in significant disparities in the charging states and energy levels of different EVs at any given time, potentially leading to perceived unfairness or dissatisfaction among EV owners.
2. Although many methodologies consider user satisfaction for EVs, these metrics are often premised upon predetermined EV connection and disconnection times. In reality, the exact disconnection time for EVs is still being determined. Relying on such methods can, at times, lead to excessive discharging of EVs, rendering them with critically low energy levels. Should owners need to use their vehicles at such moments, the residual energy may be insufficient for their travel needs.
3. Existing proposed methods effectively utilize EVs within microgrids for rational charging and discharging, thus reducing operational costs. However, they undeniably extend the charging and billing duration for EVs. Consequently, this significantly diminishes the enthusiasm and satisfaction of EV users participating in microgrid scheduling. Therefore, the development of a user compensation mechanism specifically addressing EV participation in microgrid scheduling is both crucial and necessary. Unfortunately, the existing literature on EV scheduling rarely delves into the economic compensation aspects related to their involvement in microgrid scheduling.
4. When optimizing operational methods for practical applications, environmental parameters of microgrids and the operating status of each EV are often difficult to predict. Therefore, fixed microgrid operational strategies calculated based on predicted data often fail to achieve the expected results in practical applications. This requires optimization operational methods to have higher robustness, which means that the method can autonomously adjust the output action strategy when facing different environmental conditions and EV operating statuses.

To address these issues, this paper introduces a two-stage, dual-layer optimized scheduling approach. Firstly, the charging process of EVs is divided into two stages based on their SOC: a constrained scheduling stage when their SOC is between 0 and 0.8 and a free scheduling stage when their SOC is between 0.8 and 1. In the constrained scheduling stage, EVs are only allowed to charge to meet the owners' charging needs; however, they can participate in grid scheduling by adjusting the charging power levels. During the free scheduling stage, EVs can participate in grid scheduling through charge–discharge processes, but their SOC is required to stay below 0.8. Subsequently, a two-layer optimization operational approach is proposed. The lower layer ensures uniform charging states for EVs in the constrained scheduling stage and uniform SOC levels for those in the free scheduling stage. The upper layer calculates the optimal strategy for microgrid operations using multiagent reinforcement learning, providing control reference signals for the lower layer. Notably, this article stipulates that within microgrids, only the energy consumed by EVs during the constrained scheduling stage incurs charges. Conversely, during the free scheduling stage, microgrids do not levy fees for EV charging and discharging. This approach significantly reduces the billing duration for EVs compared to traditional methods, thereby enhancing user satisfaction with EV participation in microgrid scheduling. The main research contributions of this paper are as follows:

1. A two-stage control model for EVs within a microgrid is established. This model initially categorizes the charging process of EVs into a constrained scheduling stage and a free scheduling stage based on their SOC. During the constrained scheduling stage, the EVs participate in the microgrid scheduling by adjusting their charging power.

While in the free scheduling stage, the EVs are involved in microgrid scheduling by modulating their charge–discharge rates.

2. A two-layer optimized control architecture is introduced. The lower layer is a distributed control layer, ensuring the uniform state of charging (SOC) for each EV through consensus control. The upper layer is an optimization scheduling layer, employing multiagent reinforcement learning to realize the optimal operational strategy for the microgrid, thus minimizing operational costs.
3. A novel consensus control method is designed for the lower layer. This method ensures consistent charging states for all EVs in the constrained scheduling stage and consistent SOC states for those in the free scheduling stage. Moreover, it provides a smooth transition from the constrained scheduling stage to the free scheduling stage. In addition, this method also guarantees that only a subset of EVs needs to receive upper-layer signals to control all EVs.
4. A novel two-stage MAPPO algorithm is presented for the upper layer. In this algorithm, a novel MAPPO pretraining approach is introduced. By combining pretraining with the training stage, the computational speed and effectiveness of the algorithm are significantly enhanced.

The remaining sections of this paper are structured as follows: Section 2 establishes the microgrid model with EVs; Section 3 details the proposed two-layer, two-stage operational optimization method for EV-integrated microgrids; Section 4 presents simulation analyses of the proposed method; and Section 5 offers a comprehensive conclusion.

## 2. Model of Microgrid with EVs

A salient characteristic of microgrids is their incorporation of a diverse array of energy sources to ensure a consistent supply of power [22]. The microgrid designed in this study includes microturbines (MTs), energy storage (ES), photovoltaic power generation (PV), and wind turbine power generation (WT) in addition to EVs. The microgrid is divided into two areas: the power generation resource area and the user load area. The power generation resource area contains PV, WT, MTs, and ES. The user load area contains the microgrid's power load and EV charging stations. The energy forms that can be scheduled in this microgrid include MTs, ES, and EVs. This paper adopts two deep reinforcement learning agents to compute and provide scheduling reference signals for these schedulable forms of energy. One deep reinforcement learning agent is located in the power generation resource area and provides scheduling reference signals for MTs and ES. We call it the distributed power generation agent (DG agent). Another deep reinforcement learning agent is located in the user load area and provides scheduling reference signals for EVs. We call it the EV agent. In summary, the composition of the microgrid designed in this paper is shown in Figure 1.

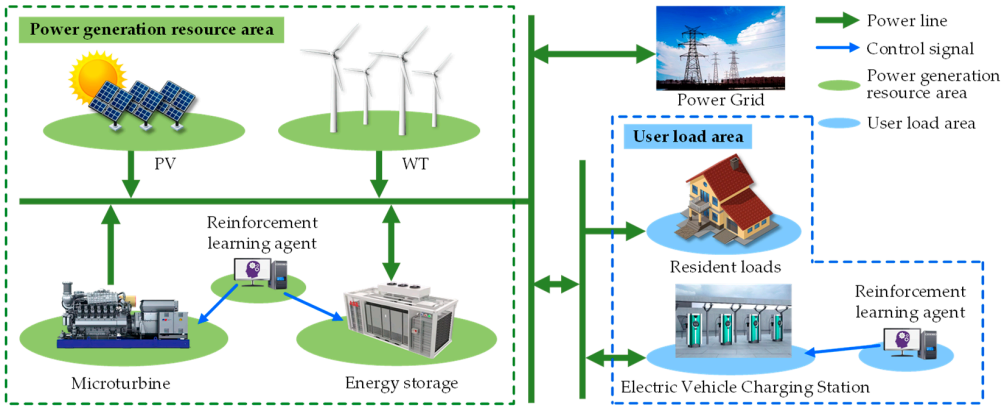


**Figure 1.** The composition structure of microgrid with EVs proposed in this paper.

Firstly, we model the EVs, MTs, ES, and the microgrid bus within the microgrid.

### 2.1. EVs Model

EVs can serve as an ES medium, participating in the energy scheduling of the microgrid. However, unlike traditional ES systems, the storage capacity of EVs varies over time. At any given moment, the storage capacity of EVs is contingent upon the number of EVs connected to the microgrid. The more EVs that are connected, the larger the schedulable storage capacity. Moreover, the charging and discharging processes of EVs need to take into consideration the electricity requirements of the EV owners. The charging and discharging procedure of each EV can be expressed as follows:

$$SOC_{\mathrm{EV},\,i}(t+1) = \begin{cases} SOC_{\mathrm{EV},\,i}(t) + \frac{\eta_{\mathrm{EV}} P_{\mathrm{EV},\,i}(t)\Delta t}{E_{\mathrm{EV},i}} & P_{\mathrm{ES},\,i}(t) > 0 \\ SOC_{\mathrm{EV},\,i}(t) + \frac{P_{\mathrm{EV},\,i}(t)\Delta t}{\zeta_{\mathrm{EV}} E_{\mathrm{EV},i}} & P_{\mathrm{ES},\,i}(t) \le 0 \end{cases} \tag{1}$$

where $SOC_{\mathrm{EV},\,i}(t)$ represents the state of charge of the $i$-th EV at time $t$; $P_{\mathrm{EV},\,i}(t)$ denotes the charging/discharging power of the $i$-th EV during time interval $t$; $E_{\mathrm{EV},i}$ signifies the battery capacity of the $i$-th EV; and $\eta_{\mathrm{EV}}$ and $\zeta_{\mathrm{EV}}$ are, respectively, the charging and discharging efficiencies of the EV.

The charging and discharging of the EVs are subject to the following constraints:

$$\begin{cases} P^{\min}_{\mathrm{EV},\,i} < |P_{\mathrm{EV},\,i}(t)| < P^{\max}_{\mathrm{EV},\,i} & P_{\mathrm{EV},\,i}(t) \neq 0 \\ P_{\mathrm{EV},\,i}(t) = 0 & P_{\mathrm{EV},\,i}(t) = 0 \end{cases} \tag{2}$$

$$SOC^{\min}_{\mathrm{EV},\,i} < SOC_{\mathrm{EV},\,i}(t) < SOC^{\max}_{\mathrm{EV},\,i} \tag{3}$$

where $P^{ch.\max}_{\mathrm{EV},\,i}$ and $P^{ch.\min}_{\mathrm{EV},\,i}$ represent the maximum and minimum charging power, respectively, for the $i$-th EV; $P^{dis.\max}_{\mathrm{EV},\,i}$ and $P^{dis.\min}_{\mathrm{EV},\,i}$, respectively, denote the maximum and minimum discharging power for the $i$-th EV; and $SOC^{\max}_{\mathrm{EV},\,i}$ and $SOC^{\min}_{\mathrm{EV},\,i}$ are the maximum and minimum SOC, respectively, for the $i$-th EV. In this study, the costs associated with EVs are limited to the operational costs of the EV charging stations. The costs generated by the EVs themselves are borne by the EV owners. It is generally accepted that the operating costs of EV charging stations comprise staff wages and equipment maintenance costs [23]. This paper assumes a constant operational cost, $C_{\mathrm{EV}}(t)$, for the EVs in each time interval.

### 2.2. MT Model

MTs provide an adjustable power supply to the microgrid by combusting fossil fuels, effectively reducing the microgrid's dependency on the external grid. Its fuel cost can be represented by a quadratic function, as shown in Equation (4).

$$C_{\mathrm{MT}}(t) = a_{\mathrm{MT}}(P_{\mathrm{MT}}(t))^2 + b_{\mathrm{MT}} P_{\mathrm{MT}}(t) + c_{\mathrm{MT}} \tag{4}$$

where $C_{\mathrm{MT}}(t)$ denotes the fuel cost for the MT during time interval $t$; $P_{\mathrm{MT}}(t)$ represents the electrical power output (in kW) of the microturbine during time interval $t$; and $a_{\mathrm{MT}}$, $b_{\mathrm{MT}}$, and $c_{\mathrm{MT}}$ are the fuel cost coefficients for the MT.

The output power of the MT is subject to the following constraints:

$$P^{\min}_{\mathrm{MT}} < P_{\mathrm{MT}}(t) < P^{\max}_{\mathrm{MT}} \tag{5}$$

$$-R_{\mathrm{MT,down}} \le P_{\mathrm{MT}}(t) - P_{\mathrm{MT}}(t-1) \le R_{\mathrm{MT,up}} \tag{6}$$

where $P^{\min}_{\mathrm{MT}}$ and $P^{\max}_{\mathrm{MT}}$ represent the maximum and minimum output power, respectively, of the MT, while $R_{\mathrm{MT,down}}$ and $R_{\mathrm{MT,up}}$ denote the upward ramp and downward ramp constraints of the MT, respectively.

### *2.3. ES Model*

The ES is composed of batteries. It can operate in harmony with renewable energy sources, which possess inherent variability and unpredictability, thus playing a role in "peak shaving and valley filling." This ensures both the reliability and economic viability of the microgrid. Taking into account the charging and discharging power of the batteries, as well as the SOC of the ES, the charging and discharging expressions for the ES are

$$SOC_{\mathrm{ES}}(t+1) = \begin{cases} SOC_{\mathrm{ES}}(t) + \frac{\eta_{\mathrm{ES}} P_{\mathrm{ES}}(t)\Delta t}{E_{\mathrm{ES}}} & P_{\mathrm{ES}}(t) > 0 \\ SOC_{\mathrm{ES}}(t) + \frac{P_{\mathrm{ES}}(t)\Delta t}{\zeta_{\mathrm{ES}} E_{\mathrm{ES}}} & P_{\mathrm{ES}}(t) \le 0 \end{cases} \tag{7}$$

where $SOC_{\mathrm{ES}}(t)$ represents the SOC of the ES at time $t$; $P_{\mathrm{ES}}(t)$ denotes the electrical power either outputted or absorbed by the ES during time interval $t$; $\eta_{\mathrm{ES}}$ and $\zeta_{\mathrm{ES}}$ are the charging and discharging efficiencies, respectively, of the ES; and $E_{\mathrm{ES}}$ signifies the rated capacity of the ES.

The cost of the ES is constituted by both capacity cost and power cost, as elaborated below:

$$C_{\mathrm{ES}}(t) = g_E E_{\mathrm{ES}} + g_P |P_{\mathrm{ES}}(t)| \Delta t \tag{8}$$

where $C_{\mathrm{ES}}(t)$ denotes the cost for the ES during time interval $t$; $g_E$ represents the capacity coefficient for the ES capacity cost; and $g_P$ is the power coefficient for the ES power cost.

The charging and discharging of the ES are subject to the following constraints:

$$\begin{cases} 0 < P_{\mathrm{ES}}(t) < P_{\mathrm{ES}}^{ch.\max} & P_{\mathrm{ES}}(t) > 0 \\ P_{\mathrm{ES}}^{dis.\max} < P_{\mathrm{ES}}(t) < 0 & P_{\mathrm{ES}}(t) \le 0 \end{cases} \tag{9}$$

$$SOC_{\mathrm{ES}}^{\min} < SOC_{\mathrm{ES}}(t) < SOC_{\mathrm{ES}}^{\max} \tag{10}$$

where $P_{\mathrm{ES}}^{ch.\max}$ and $P_{\mathrm{ES}}^{dis.\max}$, respectively, represent the maximum charging and discharging power of the ES and $SOC_{\mathrm{ES}}^{\max}$ and $SOC_{\mathrm{ES}}^{\min}$ represent the maximum and minimum states of charge, respectively, for the ES.

### *2.4. Microgrid Bus Model*

To ensure the complete absorption of renewable energy, it is assumed that all WT and PV in the microgrid are microgrid-connected. The microgrid encompasses various forms of energy, and power balance must be maintained at the microgrid bus. Assuming there are $m$ EVs in the microgrid, the power balance relationship at the microgrid bus can be expressed as

$$P_{\mathrm{grid}}(t) + \sum_{i=1}^{m} P_{\mathrm{EV},\,i}(t) + P_{\mathrm{MT}}(t) + P_{\mathrm{PV}}(t) + P_{\mathrm{WT}}(t) + P_{\mathrm{ES}}(t) = P_{\mathrm{L}}(t) \tag{11}$$

where $P_{\mathrm{grid}}(t)$ represents the power exchanged between the microgrid and the external grid; $P_{\mathrm{PV}}(t)$ and $P_{\mathrm{WT}}(t)$ denote the output power from PV and WT, respectively; and $P_{\mathrm{L}}(t)$ signifies the power consumed by the loads within the microgrid.

The cost incurred by the microgrid when purchasing electricity from the external grid, or the revenue earned from selling electricity, can be expressed as:

$$\begin{cases} C_{\mathrm{grid}}(t) = \sigma^b(t) P_{\mathrm{grid}}(t) & P_{\mathrm{grid}}(t) > 0 \\ C_{\mathrm{grid}}(t) = \sigma^s(t) P_{\mathrm{grid}}(t) & P_{\mathrm{grid}}(t) \le 0 \end{cases} \tag{12}$$

where $C_{\mathrm{grid}}(t)$ denotes the cost associated with the energy exchange between the microgrid and the external grid and $\sigma^b(t)$ and $\sigma^s(t)$, respectively, represent the electricity prices for the microgrid when purchasing from and selling to the external grid. Therefore, the total operational cost for the microgrid as proposed in this paper can be expressed as

$$F(t) = C_{\mathrm{EV}}(t) + C_{\mathrm{MT}}(t) + C_{\mathrm{ES}}(t) + C_{\mathrm{grid}}(t) \tag{13}$$

where $F(t)$ represents the total operational cost of the microgrid.

## 3. The Two-Stage, Dual-Layer Distributed Optimization Operational Approach

### *3.1. Framework of the Two-Stage, Dual-Layer Optimization Operation Approach*

In practical, everyday life, domestic EVs are predominantly used for intracity transportation. According to [24], approximately 88% of urban EVs do not exceed a daily distance of 55 km, and over 95% do not travel more than 100 km daily. The energy consumption of an EV being is around 15 kWh per 100 km [25], which implies that for an EV with a battery capacity of more than 20 kWh, reaching a 0.8 SOC level would sufficiently cater to the daily round-trip energy requirements of most domestic EVs. Given this backdrop, to fully respect EV owners' charging demands while maintaining EVs' ability to participate in grid scheduling, we divide the EV charging process into two stages. When the SOC of EVs is between 0 and 0.8, this stage is called the "Constrained Scheduling Stage". In this stage, we ensure the vehicle remains charging, but its participation in microgrid scheduling is possible by adjusting the charging power. When the battery charge is between 0.8 and 1, it is called the "Free Scheduling Stage." Here, the EV is treated as a storage battery, contributing to the microgrid's scheduling via its charging and discharging operations. It is noted that the SOC of EVs cannot fall below 0.8 at this stage.

Given the above context, we propose a two-layer distributed optimization operation method. The lower layer is the distributed control layer, where we introduce a novel consensus control method. This method ensures that EVs in the constrained scheduling stage are allocated charging power proportionally to their capacity, while those in the free scheduling stage maintain consistent SOC. The upper layer is the optimization scheduling layer, which employs multiagent reinforcement learning to calculate the optimal operation strategy of the microgrid. The advantage of this two-layer approach is that it negates the need to calculate charging and discharging strategies for individual vehicles. Instead, all EVs are considered collectively for scheduling. Specifically, the upper layer leverages deep reinforcement learning agents to provide overall EV scheduling strategies, while the lower layer employs the consistency control method to automatically translate the upper layer's strategies into respective charging or discharging reference signals per vehicle. Furthermore, as the lower layer's consensus control ensures consistency in the states of all EVs, it circumvents any contradictions that might arise due to variations in individual EV charging/discharging patterns.

For the constrained scheduling stage, we designate the reference signal as the charging power of the EV. Given the limited range of SOC variation during the free scheduling stage (0.8–1), excessive charging or discharging power could lead the EV's SOC to exceed the stipulated range. Conversely, too little charging or discharging power could potentially harm the EV's battery. Therefore, for the free scheduling stage, we designate the reference signal as the SOC value the EV needs to achieve at the end of each scheduling timestep. By setting these SOC reference values between 0.8 and 1, we ensure that the EV's charge neither falls below 0.8 nor surpasses its maximum capacity. Moreover, these SOC reference values are set as a series of discrete numbers, with the difference between two adjacent numbers guaranteeing that the EV's charging or discharging power will not fall below its minimum power threshold.

Based on the above analysis, the objective of the lower layer is to make the EV's charging power follow the power reference signal provided by the upper layer when the EV's SOC is between 0 and 0.8. When the EV's SOC exceeds 0.8, the goal is to ensure a linear increase in the EV's SOC and reach the control reference value by the end of each scheduling timestep. Meanwhile, the upper layer's objective is to calculate the optimal operating strategy for the microgrid through interactions between the EV agent and the DG agent. Furthermore, the upper layer provides control reference signals for the microgrid's EVs, MTs, and ES. The framework of the proposed two-stage, dual-layer optimization operation approach is depicted in Figure 2.

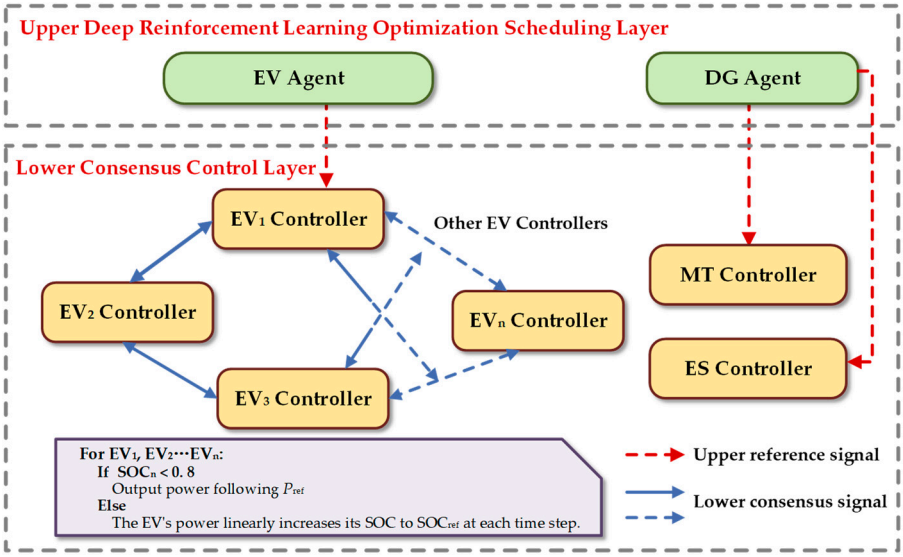


**Figure 2.** The framework of the proposed two-stage, dual-layer optimization operation approach.

### *3.2. Lower-Layer Consensus Control Method*

#### 3.2.1. Consensus Control Basics

Each controller of the EV in the microgrid can be regarded as a consensus agent, and the communication relationships among multiple consensus controllers can be represented by graph $G_u(v_u, \psi_u, K_u, K_u^0)$. Assuming there are $n_u$ agents in the graph, $v_u = \{v_{u,1} \cdots, v_{u,n_u}\}$ represents the set of nodes, each of which represents a consensus agent. $\psi_u \in v_u \times v_u$ represents the set of edges, representing the communication lines between nodes. $K_u = (K_{ij}^u)_{(n_u-1)\times(n_u-1)}$ represents the weights of edges. If there is a communication connection between $v_{u,i} \in v_u$ and $v_{u,j} \in v_u$, then $k_{ij}^u > 0$, otherwise, $k_{ij}^u = 0$. $K_u^0 = \mathrm{diag}(k_{1,0}^u, \cdots, k_{n_u,0}^u)$ represents the leading adjacency matrix. If $v_{u,i} \in v_u$ can receive a reference signal, then $k_{i0}^u > 0$, otherwise, $k_{i0}^u = 0$. Assuming that each node has a scalar state signal $x_i$, each node can update its state based on its own state and the state signal of the nodes it communicates with. Based on the consensus control scheme, the rules for updating the state of the node can be expressed as follows [26]:

$$\dot{x}_i(t) = \sum_{j \in v_u} [k_{ij}^u (x_j(t) - x_i(t)) + k_{i0}^u (x_{ref} - x_i(t))] \tag{14}$$

where $\dot{x}_i$ denotes the differential of the state variable $x_i$. According to [26], if the communication network graph among consensus agents has a spanning tree, then the following theorem holds.

**Theorem 1.** *If the update rule defined by (13) is employed, then the states of all nodes will converge to the reference value $x_{ref}$, i.e.,*

$$\lim_{t \to \infty} x_i(t) = x_{ref} \tag{15}$$

The proof process of the theorem mentioned above can be found in [26]. Notably, the reference value $x_{ref}$ can also possess dynamics.

#### 3.2.2. Lower-Layer Two-Stage Consistency Control Method

Based on the analysis in Section 3.1 the control objectives for the lower layer can be formulated as follows: for all $i$ and $j$ less than $m$, we have the following.

When $0 < SOC(t) < 0.8$, it is in the constrained scheduling stage. One of our aims is for the charging power of the EVs to be allocated in proportion to their capacity, i.e.,

$$\lim_{t\to\infty} |P_{\mathrm{EV},\,i}(t)/P^{\max}_{\mathrm{EV},\,i} - P_{\mathrm{EV},\,j}(t)/P^{\max}_{\mathrm{EV},\,j}| = 0 \tag{16}$$

We set the capacity ratio $\eta = P^{\max}_{\mathrm{EV},\,i}/P_{\mathrm{nom}}$, where $P_{\mathrm{nom}}$ is a constant value. Therefore, Equation (16) can be expressed as

$$\lim_{t\to\infty} |P_{\mathrm{EV},\,i}(t)/\eta - P_{\mathrm{EV},\,j}(t)/\eta| = 0 \tag{17}$$

In order to facilitate representation, we let $P_{k,i}(t) = P_{\mathrm{EV},\,i}(t)/\eta$ and treat $P_{k,i}(t)$ as the object for consensus control. We aim for $P_{k,i}(t)$ to follow the control reference signal provided by the upper layer, that is:

$$\lim_{t\to\infty} |P_{k,i}(t) - P_{ref}(t)| = 0 \tag{18}$$

where $P_{ref}$ is the power reference signal of the upper-layer EV agent. According to (13), when $0 < \mathrm{SOC} < 0.8$, the control expression of the EV can be written as

$$\dot{P}_{k,i}(t) = \sum_{j=1}^{m} [k^P_{ij}(P_{k,j}(t) - P_{k,i}(t)) + k^P_{i0}(P_{ref}(t) - P_{k,i}(t))] \tag{19}$$

In accordance with Theorem 1, Equation (19) ensures that the charging power of each EV is allocated proportionally to its capacity and guarantees that $P_{k,i}$ follows the power reference signal provided by the upper-layer EV agent.

When $0.8 < SOC(t) < 1$, it is in the free scheduling stage. We let the reference signal for the EV's SOC be $SOC_{k,i}$, then the control objective for consensus control can be expressed as

$$\lim_{t\to\infty} |SOC_{k,i}(t) - SOC_{k,j}(t)| = 0 \tag{20}$$

$$\lim_{t\to\infty} |SOC_{k,i}(t) - SOC_{k,ref}(t)| = 0 \tag{21}$$

In the equation, $SOC_{k,ref}$ represents the SOC reference signal provided by the upper-layer EV agent. The updated expression for $SOC_{k,i}$ can be written as

$$\dot{SOC}_{k,i}(t) = \sum_{j=1}^{m} [k^S_{ij}(SOC_{k,j}(t) - SOC_{k,i}(t)) + k^S_{i0}(SOC_{ref}(t) - SOC_{k,i}(t))] \tag{22}$$

It is worth noting that the $SOC_{k,i}(t)$ here does not represent the actual charge of the EV. Instead, it signifies the SOC that the EV should attain at the end of this scheduling timestep. Given that only a few controllers are directly connected to the upper-layer agents, based on Theorem 1, Equation (22) ensures that all EVs can attain the SOC level provided by upper layer at the end of each scheduling timestep.

As analyzed in Section 3.1, our control objective during the free scheduling stage is to linearly increase or decrease the SOC of the EV by controlling its charging and discharging power. The aim is to ensure that at the end of each scheduling timestep, the SOC level of each EV matches the reference value provided by the upper layer. Hence, we have designed the charging and discharging formula for the EV during the free scheduling stage as

$$P_{\mathrm{EV},\,i}(t) = \frac{P^{\max}_{\mathrm{EV},\,i} \cdot (SOC_{k,i}(t) - SOC_{\mathrm{EV},\,i}(t))}{T_i - t(\mathrm{mod}\ T_i) + \kappa} \tag{23}$$

where $T_i$ denotes the length of each scheduling timestep, while $t(\mathrm{mod}\ T_i)$ represents the remainder when time $t$ is divided by the scheduling timestep length $T_i$. $\kappa$ is an extremely

small positive number that is intended to prevent the divergence of results caused by a denominator of zero.

In summary, the lower-level controller designed in this study can be expressed as

$$\begin{cases} \dot{P}_{k,i}(t) = \sum\limits_{j=1}^{m} [k_{ij}^{P}(P_{k,j}(t) - P_{k,i}(t)) + k_{i0}^{P}(P_{ref}(t) - P_{k,i}(t))] & 0 < SOC(t) < 0.8 \\ P_{\mathrm{EV},\,i}(t) = P_{k,i}(t) \cdot \eta & \\ \dot{SOC}_{k,i}(t) = \sum\limits_{j=1}^{m} [k_{ij}^{S}(SOC_{k,j}(t) - SOC_{k,i}(t)) + k_{i0}^{S}(SOC_{ref}(t) - SOC_{k,i}(t))] & 0.8 < SOC(t) < 1 \\ P_{\mathrm{EV},\,i}(t) = \frac{P_{\mathrm{EV},\,i}^{\max} \cdot (SOC_{k,i}(t) - SOC_{\mathrm{EV},\,i}(t))}{T_i - t(\mathrm{mod}\ T_i) + \kappa} & \end{cases} \tag{24}$$

To enhance the convergence rate of the consensus algorithm, this study adopts the improved average consensus method as detailed in [27,28] to determine the values of $k_{ij}^{P}$, $k_{ij}^{S}$, $k_{i0}^{P}$, and $k_{i0}^{S}$ in Equation (24). Their expressions are presented as follows:

$$k_{ij}^{P} = k_{ij}^{S} = \begin{cases} \lambda / [(n_i + n_j)/2] & i-\text{th and } j-\text{th EVs have communication links.} \\ 0 & i-\text{th and } j-\text{th EVs have no communication links.} \end{cases} \tag{25}$$

$$k_{i0}^{P} = k_{i0}^{S} = \begin{cases} \lambda / [(n_i + n_0)/2] & i-\text{th EV and upper layer have communication links.} \\ 0 & i-\text{th EV and upper layer have no communication links.} \end{cases} \tag{26}$$

where $n_i$ and $n_j$, respectively, denote the number of communication links connected to the $i$ and $j$ EVs and $n_0$ represents the number of communication links connected to the upper-layer EV agent.

### *3.3. Upper-Layer Optimization Scheduling Method*

#### 3.3.1. Markov Decision Process in Microgrids with EVs

The upper-layer optimization scheduling for the microgrid proposed in this study utilizes a multiagent deep reinforcement learning approach. The decision-making process of deep reinforcement learning can be described as a Markov decision process (MDP) [29]. An MDP typically comprises five elements, namely, $\{S,\ A,\ P_{S,S'},\ r,\ \gamma\}$. Specifically, $S$ represents the state space, which is the set of environment-state information observable by the agent; $A$ denotes the action space, signifying the set of actions that the agent can undertake; $P_{s,s'}$ indicates the state transition probability, representing the probability that the environment transitions from state $S$ to state $S'$ when the agent takes action $a$; $r$ is the immediate reward, signifying the immediate reward the environment gives to the agent upon taking action $a$ in state $S$; and $\gamma$ is the discount factor, depicting the influence of the current action on the rewards obtained by the agent in future timesteps. For the microgrid discussed in this paper, its state space, action space, and reward function can be designed as follows:

- State space

The state space refers to the set of environmental information observable by the deep reinforcement learning agent. The state space of the EV-inclusive microgrid designed in this study encompasses operational time, user load, WT power, PV power, total EVs power, microgrid bus power, MT power, ES power, ES SOC, and external grid time-of-use electricity prices. EVs in the microgrid are controlled by the EV agent, whereas MTs and ES are controlled by the DG agent. As the EV and DG agents operate in different areas, the environmental state variables they can observe differ. Variables observable by the EV agent include system operational time, total EV power, user load, microgrid bus power, and external grid time-of-use prices. Variables observable by the DG agent include system operational time, PV power, WT power, MT power, ES power, and ES SOC. Thus, the state spaces for the EV and DG agents can be separately described as

$$s_{\mathrm{EV},t} = [t,\ P_{\mathrm{EV,sum}}(t),\ P_{\mathrm{L}}(t),\ P_{\mathrm{grid}}(t),\ \sigma^{b}(t),\ \sigma^{s}(t)] \tag{27}$$

$$s_{\mathrm{DG},t} = [t,\ P_{\mathrm{PV}}(t),\ P_{\mathrm{WT}}(t),\ P_{\mathrm{MT}}(t),\ P_{\mathrm{ES}}(t),\ SOC_{\mathrm{ES}}(t)] \tag{28}$$

where $S_{\mathrm{EV},t}$ denotes the state space of the EV agent; $S_{\mathrm{DG},t}$ represents the state space of the DG agent; and $P_{\mathrm{EV},\mathrm{sum}}$ indicates the total power of the EVs, which can be acquired from the microgrid bus connected to the EV charging station.

- Action space

In the microgrid designed for this study, the actions output by the EV agent encompass the power reference signal and the SOC reference signal for EVs. The actions output by the DG agent comprise the power reference signals for both MTs and ES. Consequently, the action spaces for the EV and DG agents can be, respectively, defined as

$$a_{\mathrm{EV},t} = [P_{ref}(t),\ SOC_{ref}(t)] \tag{29}$$

$$a_{\mathrm{DG},t} = [P_{\mathrm{MT},\,ref}(t),\ P_{\mathrm{ES},\,ref}(t)] \tag{30}$$

where $a_{\mathrm{EV},t}$ denotes the action space of the EV agent at time $t$; $a_{\mathrm{DG},t}$ represents the action space of the DG agent at time $t$; and $P_{\mathrm{MT},\,ref}(t)$ and $P_{\mathrm{ES},\,ref}(t)$ respectively indicate the power reference signals for MTs and ES.

- Reward function

After the selection of any action by the deep reinforcement learning agent, the environment provides a reward. However, if the chosen action leads the microgrid to operate outside the environmental constraints, a penalty is given by the environment. In this study, the environmental constraint penalties arise from the ramping constraints of MT and the SOC constraints of the energy storage. The penalty expressions are, respectively:

$$C^{c}_{\mathrm{MT}}(t) = \lambda^{c}_{\mathrm{MT}} \cdot \max\{P_{\mathrm{MT}}(t) - P_{\mathrm{MT}}(t-1) - R_{\mathrm{up}},\ 0\} - \lambda^{c}_{\mathrm{MT}}\min\{P_{\mathrm{MT}}(t) - P_{\mathrm{MT}}(t-1) + R_{\mathrm{down}},\ 0\} \tag{31}$$

$$C^{S}_{\mathrm{ES}}(t) = \lambda^{S}_{\mathrm{ES}} \cdot E_{\mathrm{ES}}\max\{SOC_{\mathrm{ES}}(t) - SOC^{\max}_{\mathrm{ES}},\ 0\} - \lambda^{S}_{\mathrm{ES}} \cdot E_{\mathrm{ES}}\min\left\{SOC_{\mathrm{ES}}(t) - SOC^{\min}_{\mathrm{ES}},\ 0\right\} \tag{32}$$

where $C^{c}_{\mathrm{MT}}$ denotes the penalty when the difference in output power between two consecutive time instants for MT exceeds its ramping constraints. $\lambda^{c}_{\mathrm{MT}}$ represents the penalty coefficient for the MT's ramping constraints. $C^{S}_{\mathrm{ES}}(t)$ signifies the penalty when the SOC of the ES exceeds its constraints. $\lambda^{S}_{\mathrm{ES}}$ is the penalty coefficient for the energy storage's SOC constraints.

For ES, the SOC at the final moment of the current scheduling period is taken as the SOC at the beginning of the next scheduling period. In order to ensure that the SOC at the end of the current scheduling period does not impact the scheduling ability of the energy storage in the following period, we desire the SOC at the conclusion of each scheduling period to be as close as possible to its initial value. Therefore, we have designed an exponential form for the ES's SOC reset penalty:

$$C^{r}_{\mathrm{ES}}(t) = \lambda^{r}_{\mathrm{ES}} \cdot (e^{\delta \cdot t} - 1) \cdot [SOC(t) - SOC(0)]^2 \tag{33}$$

where $C^{r}_{\mathrm{ES}}$ denotes the ES's reset penalty; $\lambda^{r}_{\mathrm{ES}}$ represents the penalty coefficient for the reset penalty; and $\delta$ signifies the exponential coefficient for the reset penalty. Within a scheduling period, the initial reset penalty for ES is minimal. As the scheduling time progresses, the reset penalty for energy storage will increase, reaching its maximum at the end of the scheduling period.

In summary, the upper-layer optimization scheduling primarily focuses on economic considerations. It aims to minimize the operational costs of the microgrid within a schedul-

ing period by judiciously controlling the EVs, MTs, and ES. Thus, the total reward function of the EV and DG agents can be expressed as

$$R = -\sum_{t=1}^{T} [F(t) + C_{\mathrm{MT}}^{c}(t) + C_{\mathrm{ES}}^{S}(t) + C_{\mathrm{ES}}^{r}(t)] \tag{34}$$

where $R$ represents the total reward of the EV and DG agents.

3.3.2. Proximal Policy Optimization Algorithm

Proximal policy optimization (PPO) is an on-policy deep reinforcement learning method developed by OpenAI in 2017, and it serves as the default deep reinforcement learning algorithm utilized by OpenAI [30]. Compared with off-policy deep reinforcement learning algorithms like deep Q-network (DQN) and deep deterministic policy gradient (DDPG), the PPO algorithm typically exhibits superior stability and convergence. The PPO algorithm is generally composed of one critic network and two actor networks. The training process of the PPO algorithm can be divided into three stages: data collection, data processing, and network training, as illustrated in Figure 3.

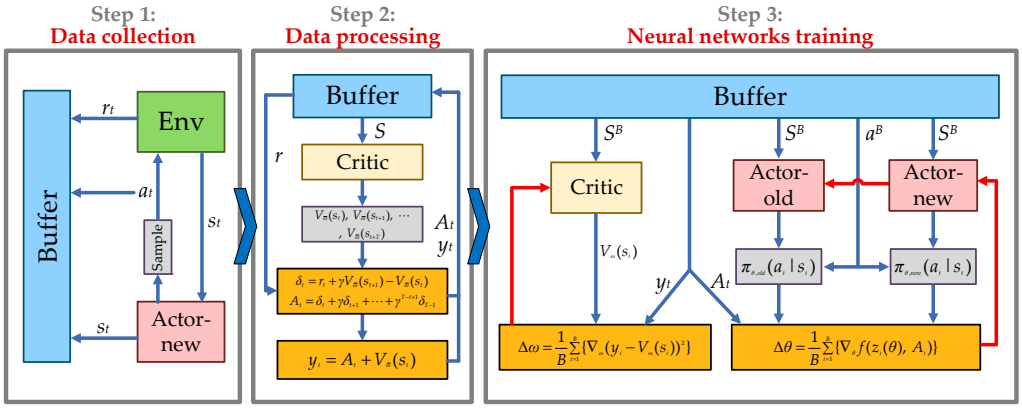


**Figure 3.** The training process of the PPO algorithm.

As depicted in Figure 3, we have $S = \{s_t, s_{t+1}, \cdots, s_{t+T-1}\}$, $a = \{a_t, a_{t+1}, \cdots, a_{t+T-1}\}$, $r = \{r_t, r_{t+1}, \cdots, r_{t+T-1}\}$, $S^B = \{s_1, s_2, \cdots, s_B\}$, and $a^B = \{a_1, a_2, \cdots, a_B\}$. During the data collection stage, the agent's actor network outputs a probability distribution of various actions based on the environmental state $s_t$. Subsequent action $a_t$ is generated through probability sampling. The environment then provides the immediate reward $r_t$ for action $a_t$ under state $s_t$. Next, the agent stores the environmental state $S$, action $a$, and immediate reward $r$ into the experience replay buffer D. It is important to note that during this data collection stage, only the data from the agent's interaction with the environment is stored in D, and there is no update occurring within the agent's neural networks. Once the replay buffer is filled, the data collection stage ends, and the data processing stage begins.

During the data processing stage, the PPO's critic network initially generates an evaluation $V_{\overline{\omega}}(s_t)$ for each state $s_t$ based on the states in D. Here, $\overline{\omega}$ represents the neural network parameters of the critic network during the data processing stage. It is noteworthy that the critic network does not undergo any parameter updates in this stage; thus, $\overline{\omega}$ remains constant. Subsequently, immediate rewards $r$ for all timesteps are retrieved from D. The advantage estimation $A_t$ for each timestep is then derived using the following two equations [31]:

$$\delta_t = r_t + \gamma V_{\overline{\omega}}(s_{t+1}) - V_{\overline{\omega}}(s_t) \tag{35}$$

$$A_t = \delta_t + \gamma\delta_{t+1} + \cdots + \cdots + \gamma^{T-t+1}\delta_{T-1} \tag{36}$$

After obtaining the advantage estimation $A_t$ for each step, the target value for evaluations across all timesteps, denoted as $y_t$, is computed using the subsequent equation:

$$y_t = A_t + V_{\overline{\omega}}(s_t) \tag{37}$$

In the final step, the obtained $A_t$ and $y_t$ are stored in D to form the dataset $\{s_t,\ a_t,\ r_t,\ A_t,\ y_t\}_{t=1}^{T}$. Subsequently, the data sequences within the replay buffer are randomized, transitioning into the network training stage.

During the network training stage, the PPO algorithm employs two actor networks. One of these is used for decision-making interactions with the environment post-training; this is termed the "actor-new" network. The other is used to regulate the magnitude of updates to the actor-new network, preventing excessive updates that could destabilize the training process. As this latter actor network remains static during the update of the actor-new network and only assimilates the updated parameters from the actor-new network after its update, it is termed the "actor-old" network.

During the training process, the agent sequentially retrieves $B$ batches of data from the beginning of D and reindexes these data as $\{s_i,\ a_i,\ r_i,\ A_i,\ y_i\}_{i=1}^{B}$. Subsequently, all the action data $a_i$ from these $B$ batches are inputted simultaneously into both the actor-new and actor-old networks. Each actor network then produces the probability distributions $\pi_{\theta,new}(a_t|s_t)$ and $\pi_{\theta,old}(a_t|s_t)$ for potential output actions under each state $s_i$. The policy gradient $\Delta\theta$ for the parameters of the actor-new network, $\theta$, is then computed using the following formula:

$$z_t(\theta) = \frac{\pi_{\theta,new}(a_t|s_t)}{\pi_{\theta,old}(a_t|s_t)} \tag{38}$$

$$f(z_t(\theta),\ A_t) = \min(z_t(\theta)A_t,\ \mathrm{clip}(z_t(\theta), 1-\varepsilon,\ 1+\varepsilon)A_t) \tag{39}$$

$$\Delta\theta = \frac{1}{B}\sum_{i=1}^{B}\{\nabla_\theta f(z_i(\theta),\ A_i)\} \tag{40}$$

In (38)

$$\mathrm{clip}(z_t(\theta), 1-\varepsilon,\ 1+\varepsilon) = \begin{cases} z_t(\theta) & 1-\varepsilon \le z_t(\theta) \le 1+\varepsilon \\ 1-\varepsilon & z_t(\theta) < 1-\varepsilon \\ 1+\varepsilon & z_t(\theta) > 1-\varepsilon \end{cases} \tag{41}$$

where $\nabla_\theta f(\cdot)$ represents the gradient of the function $f(\cdot)$ with respect to the parameter $\theta$. $\varepsilon$ is a positive number between 0 and 1. Subsequently, with the aim of maximizing the policy gradient $\Delta\theta$, the gradient ascent method is employed to update the parameters of the actor-new network. Concurrent with the update of the actor network, all $s_i$ from the $B$ batches are fed into the critic network, which then produces the value estimate $V_\omega(s_i)$ for each state $s_i$. The policy gradient $\Delta\omega$ for the critic network's parameters $\omega$ is then determined using the following equation:

$$\Delta\omega = \frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_\omega(y_i - V_\omega(s_i))^2\right\} \tag{42}$$

Subsequently, with the objective of minimizing the policy gradient $\Delta\omega$, the gradient descent method is employed to update the parameters of the critic network, thus completing one training iteration of neural network training within the PPO algorithm. In the next training iteration, data are fetched from the second dataset onwards in sets of $B$ batches. This process of parameter updating continues in a similar fashion. Once the data from these $B$ batches extend to the final set in D and the neural network parameters are updated accordingly, one cycle of neural network training is finalized. After executing multiple training cycles, the network training stage concludes. At this juncture, the neural network parameters $\theta$ of the actor-new network are assigned to the neural network parameters $\theta_{old}$

of the actor-old network. D is then emptied, and the system reenters the data collection stage to newly acquire interaction data between the agent and the environment.

#### 3.3.3. Multiagent PPO Algorithm

To enhance the safety and flexibility of microgrid operations, the microgrid designed in this study utilizes two distinct deep reinforcement learning agents to compute the optimal scheduling strategy for the microgrid. The EV agent is responsible for emitting reference signals for EVs, while the DG agent handles the power reference signals for MT and ES. Designing separate agents to govern various energy units in the microgrid offers a key advantage: if one deep reinforcement learning agent malfunctions or incurs damage, it does not hamper the decision-making capabilities of the other agent. This structure thus bolsters the safety and flexibility of the microgrid's operations. Moreover, as distributed energy resources and EV charging stations are located om different areas within the microgrid, employing distinct agents to manage them can alleviate computational strain and diminish communication costs in the microgrid. Based on the above analysis, we adapt the centralized training and decentralized decision-making approach, expanding the PPO algorithm to a multiagent PPO (MAPPO) algorithm in the context of the microgrid environmental model discussed in this paper.

For notation convenience, the set of state information observed by both the EV and DG agents is termed as the global state information, while the sets of state information individually perceived by the EV and DG agents are referred to as their respective local state information. The training process of the multiagent PPO algorithm is characterized by the fact that during the data collection stage, the state and action quantities of both the EV and DG agents are aggregated into D, i.e., $s_t = (s_{\mathrm{EV},t},\ s_{\mathrm{DG},t})$ and $a_t = (a_{\mathrm{EV},t},\ a_{\mathrm{DG},t})$. During the network training stage, the actor networks of the two agents update based on their local state information, while the critic network updates are influenced by global state data. Using the EV agent as an example, during the data processing and network updating stages, the critic network of the EV agent derives evaluation values $V_{\overline{\omega}}(s_t)$ and $V_{\omega}(s_t)$ according to the global state information $s_t$ retrieved from D. Subsequently, the neural network parameters $\omega$ of the critic network are updated based on $V_{\overline{\omega}}(s_t)$ and $V_{\omega}(s_t)$. The actor-new and actor-old networks of the EV agent, meanwhile, generate action probabilities $\pi_{\theta,new}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})$ and $\pi_{\theta,old}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})$ based on the local observations $s_{\mathrm{EV},t}$ from D. Finally, the neural network parameters $\theta_{\mathrm{new}}$ and $\theta_{\mathrm{old}}$ for the actor-new and actor-old networks are subsequently updated based on $\pi_{\theta,new}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})$ and $\pi_{\theta,old}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})$, respectively. Consequently, the neural network parameter update formulas for the critic networks of the EV and DG agents, $\omega_{\mathrm{EV}}$ and $\omega_{\mathrm{DG}}$, are, respectively, given as

$$\Delta\omega_{\mathrm{EV}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_{\omega_{\mathrm{EV}}}\left(y_i - V_{\omega_{\mathrm{EV}}}(s_i)\right)^2\right\} \tag{43}$$

$$\Delta\omega_{\mathrm{DG}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_{\omega_{\mathrm{DG}}}\left(y_i - V_{\omega_{\mathrm{DG}}}(s_i)\right)^2\right\} \tag{44}$$

The update formulas for the neural network parameters $\omega_{\mathrm{EV}}$ and $\omega_{\mathrm{DG}}$ of the critic networks for the EV and DG agents are, respectively, given as

$$z_{\mathrm{EV},t}(\theta_{\mathrm{EV}}) = \frac{\pi_{\theta_{\mathrm{EV}},new}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})}{\pi_{\theta_{\mathrm{EV}},old}(a_{\mathrm{EV},t}|s_{\mathrm{EV},t})} \tag{45}$$

$$z_{\mathrm{DG},t}(\theta_{\mathrm{DG}}) = \frac{\pi_{\theta_{\mathrm{DG}},new}(a_{\mathrm{DG},t}|s_{\mathrm{DG},t})}{\pi_{\theta_{\mathrm{DG}},old}(a_{\mathrm{DG},t}|s_{\mathrm{DG},t})} \tag{46}$$

$$\Delta\theta_{\mathrm{EV}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_{\theta_{\mathrm{EV}}} f(z_{\mathrm{EV},i}(\theta_{\mathrm{EV}}),\ A_i)\right\} \tag{47}$$

$$\Delta\theta_{\mathrm{DG}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_{\theta_{\mathrm{DG}}} f\left(z_{\mathrm{DG},i}(\theta_{\mathrm{DG}}),\ A_i\right)\right\} \tag{48}$$

We assume that each episode of the agent's training comprises $T$ timesteps, and the training procedure is repeated $M$ times to guarantee the convergence of the algorithm. The detailed workflow of the MAPPO algorithm for the microgrid with EVs is illustrated in Algorithm 1.

**Algorithm 1:** MAPPO-Based Optimized Scheduling Method for Microgrid with EV

Initialize the neural networks and the parameter setting for the microgrid model at t = 0.
Input : environment, observation space $s_t$ and action space $a_t$.
Output: optimal scheduling strategies for EVs, energy storage, and gas turbines
for episode = 1 to M do
Reset environment (t = 0) to obtain observation $s_{\mathrm{EV},t}$ for EV agent and $s_{\mathrm{DG},t}$ for DG agent.
  for t = 1 to T do
    The actor-new network of EV agent and DG agent generates probability distributions for each action.
    EV agent and DG agent select action $a_{\mathrm{EV},t}$ and $a_{\mathrm{DG},t}$ by probability sampling, respectively
    EV agent obtain observation $s_{\mathrm{EV},t}$, action $a_{\mathrm{EV},t}$ and reward $r_t$. DG agent obtain observation $s_{\mathrm{DG},t}$, action $a_{\mathrm{DG},t}$ and reward $r_t$.
    Merge $\{s_{\mathrm{EV},t},\ a_{\mathrm{EV},t},\ r_t\}$ and $\{s_{\mathrm{DG},t},\ a_{\mathrm{DG},t},\ r_t\}$ into $\{s_t,\ a_t,\ r_t\}$ and store $\{s_t,\ a_t,\ r_t\}$ in D
  end
  Critic network compute $\{V_{\overline{\omega}}(s_t)\}_{t=1}^{T}$,
  Compute $\{A_t\}_{t=1}^{T}$ and $\{y_t\}_{t=1}^{T}$ use (36) and (37), respectively
  Gat date $\{s_t,\ a_t,\ r_t,\ A_t,\ y_t\}_{t=1}^{T}$ and store them in D
  for k = 1, 2, . . ., K
    Shuffle the data's order and renumber in D
    for $j = 0,\ 1,\ \cdots,\ T/B-1$
      Select B group of data $\{s_i,\ a_i,\ r_i,\ y_i,\ A_i\}_{t=1+Bj}^{B(j+1)}$
      Compute Calculate the gradients $\Delta\omega_{\mathrm{EV}}$, $\Delta\omega_{\mathrm{DG}}$, $\Delta\theta_{\mathrm{EV}}$, $\Delta\theta_{\mathrm{DG}}$ by Equations (43), (44), (47), and (48).
      Apply gradient descent on $\omega_{\mathrm{EV}}$, $\omega_{\mathrm{DG}}$ using $\Delta\omega_{\mathrm{EV}}$, $\Delta\omega_{\mathrm{DG}}$ by Adam
      Apply gradient ascent on $\theta_{\mathrm{EV}}$, $\theta_{\mathrm{DG}}$ using $\Delta\theta_{\mathrm{EV}}$, $\Delta\theta_{\mathrm{DG}}$ by Adam
    end
  end
  update $\theta_{\mathrm{EV},\,old} \leftarrow \theta_{\mathrm{EV}}$ and $\theta_{\mathrm{DG},\,old} \leftarrow \theta_{\mathrm{EV}}$
  Empty D
end

#### 3.3.4. Two-Stage PPO Training Approach

Although the PPO algorithm can address nonlinear optimization problems, it still tends to converge slowly and may easily settle into local optima. To address these issues, this paper proposes a two-stage PPO agent training method, as depicted in Figure 4.

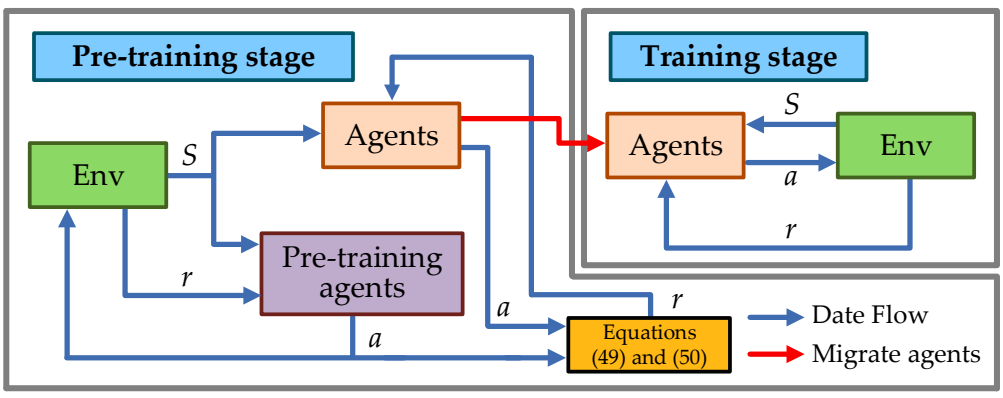


**Figure 4.** The two-stage PPO agent training method proposed in this paper.

The two-stage PPO agent training method proposed in this paper consists of a pre-training stage and a training stage, which are described in detail as follows:

- Stage 1: pre-training stage

First, a pretraining agent is prepared for the EV agent and DG agent, i.e., pretraining EV agent and pretraining DG agent, respectively. For convenience in description, we denote the pretraining EV agent and pretraining DG agent as pre-agents while defining the EV and DG agents as proto-agents. The structure of the pre-agents is identical to the proto-agents, maintaining the same dimensionality in the action space; however, the number of actions available for selection in each dimension is much less in the pre-agents. We define the action space of pre-agents as the pretraining action space and the action space of the proto-agents as the prototype action space.

During the pretraining stage, both pretraining agents and proto-agents observe state information $S$ from the environment independently and generate actions $a_t^{pre}$ and $a_t$, respectively, where $a_t^{pre} = \left\{a_{\mathrm{EV},t}^{pre}, a_{\mathrm{DG},t}^{pre}\right\}$, $a_{\mathrm{EV},t}^{pre} = [P_{ref}^{pre}(t), SOC_{ref}^{pre}(t)]$, and $a_{\mathrm{DG},t}^{pre} = [P_{\mathrm{MT},\,ref}^{pre}(t), P_{\mathrm{ES},\,ref}^{pre}(t)]$. $P_{ref}^{pre}(t)$, $SOC_{ref}^{pre}(t)$ represent the reference signal for the charging and discharging of EVs output by the pretraining EV agent. $P_{\mathrm{MT},\,ref}^{pre}(t)$, $P_{\mathrm{ES},\,ref}^{pre}(t)$ represent the power reference signals for MTs and ES output by the DG agent. Both $a_t^{pre}$ and $a_t$ are in discrete action spaces, but the number of available actions in each dimension of $a_t^{pre}$ is much lesser compared to $a_t$.

Upon receiving the action information $a_t^{pre}$ output by the pretraining agents, the environment updates its state and provides immediate rewards $R$ as per Equation (33). Concurrently, utilizing actions $a_t^{pre}$ and $a_t$, rewards for the EV and DG agents during the pretraining stage, denoted as $R_{pre}^{\mathrm{EV}}$ and $R_{pre}^{\mathrm{DG}}$, are computed using the following equation:

$$R_{pre}^{\mathrm{EV}} = -K' \sum_{i=1}^{T} \left[ \left(P_{ref}^{pre}(i) - P_{ref}(i)\right)^2 + \left(SOC_{ref}^{pre}(i) - SOC_{ref}(i)\right)^2 \right] \tag{49}$$

$$R_{pre}^{\mathrm{DG}} = -K' \sum_{i=1}^{T} \left[ \left(P_{\mathrm{MT},\,ref}^{pre}(i) - P_{\mathrm{MT},\,ref}(i)\right)^2 + \left(P_{\mathrm{ES},\,ref}^{pre}(i) - P_{\mathrm{ES},\,ref}(i)\right)^2 \right] \tag{50}$$

Ultimately, the pre-agents update their internal neural networks based on $R$, and the EV and DG agents update their internal neural networks independently based on $R_{pre}^{\mathrm{EV}}$ and $R_{pre}^{\mathrm{DG}}$, respectively. By repeating the above process multiple times, pretrained original agents are obtained.

It is noteworthy that during the pretraining stage, only the actions output by the pre-agents interact with the environment and receive rewards $R$ from the environment, aimed at finding the optimal scheduling strategy under the pretraining action space. The goal of the proto-agents is to make their output actions as close as possible to the actions output by the pre-agents. Additionally, both the EV and DG agents within the proto- and pretraining agents adopt the centralized training method depicted in Algorithm 1, while a decentralized training method is applied between the proto- and pretraining agents. Since the number of choices available to the pre-agents is small, the number of policies that can be composed is also relatively fewer, making it easier for pre-agents to find the optimal scheduling strategy. Furthermore, from Equations (48) and (49), it can be seen that the reward function of the proto-agents is linear; thus, their learning process is tantamount to imitation learning, which means the update process for the proto-agents is also relatively rapid.

- Stage 2: training stage

After the completion of the pretraining process, we extract the pretrained proto-agents. In the training stage, we allow the proto-agents to interact with the environment, output actions to the environment, and receive rewards $R$ according to Equation (33). The EV and DG agents, with the objective of maximizing reward $R$, employ the centralized

training method depicted in Algorithm 1 to learn the optimal scheduling strategy within the prototype action space.

The two-stage training approach described above is conducted in an offline simulation environment. Upon the completion of the proposed two-stage training, the MT and ES agents can then proceed to utilize their respective actor networks to make distributed online decisions in the real-world microgrid environment. In conclusion, the overall flow of the two-stage, dual-layer optimized operation method for microgrids proposed in this paper is illustrated in Figure 5.

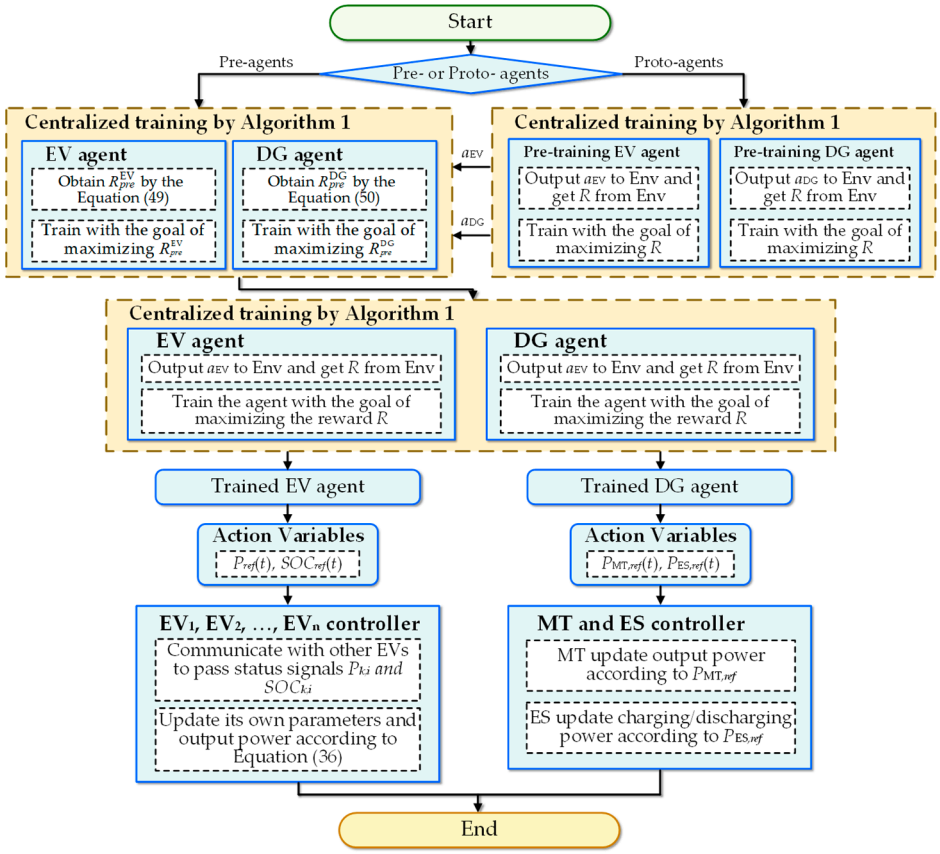


**Figure 5.** The overall flow of the two-stage, dual-layer optimized operation method for microgrids proposed in this paper.

## 4. Simulation Analysis

This paper designs a microgrid structure that incorporates PV, WT, MT, ES, and an EV charging station. We assume that the number of charging piles in the EV charging station of the microgrid is sufficient for daily EV usage. We set up 12 EV charging piles in the microgrid EV charging station, and the adjacent charging piles interact with each other through communication links to exchange status information. The spatial structure of the EV charging station is designed as shown in Figure 6. We assume that the microgrid has three types of EVs with capacities of 40 kWh, 50 kWh, and 60 kWh. The original arrival time of EVs is simulated based on the distribution of taxi shift time in Beijing [32]. The connection time of EVs is set to 8–11 h. The initial SOC levels of EVs are assumed to follow a Gaussian distribution with means and standard deviations of 60% and 10%, respectively [33]. In summary, the relevant connection information for the 12 EVs in the microgrid is shown in Table 1.

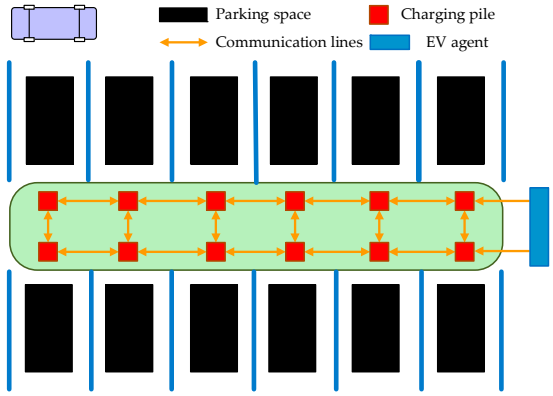


**Figure 6.** The spatial structure of the EV charging station in the microgrid proposed in this paper.

**Table 1.** Connection information for 12 EVs in the microgrid.

| EV No. | EV Battery Capacity/kW·h | Time of Arrival | Connection Duration/h | Initial SOC |
|---|---|---|---|---|
| 1 | 60 | 08:00 | 10 | 0.8 |
| 2 | 50 | 11:00 | 10 | 0.7 |
| 3 | 50 | 14:00 | 8 | 0.7 |
| 4 | 80 | 16:00 | 9 | 0.8 |
| 5 | 60 | 16:00 | 10 | 0.9 |
| 6 | 60 | 16:00 | 8 | 0.8 |
| 7 | 50 | 18:00 | 9 | 0.6 |
| 8 | 80 | 18:00 | 10 | 0.7 |
| 9 | 60 | 18:00 | 10 | 0.8 |
| 10 | 60 | 20:00 | 10 | 0.6 |
| 11 | 80 | 20:00 | 11 | 0.4 |
| 12 | 60 | 23:00 | 8 | 0.5 |

The PV power, WT power, and load data within the microgrid are depicted in Figure 7. The electricity prices for buying and selling from the external grid in the microgrid are illustrated in Table 2 [34]. The key parameters of the major equipment within the microgrid are presented in Table 3. The state spaces of the pre-agents and the proto-agents are identical, with the upper and lower bounds of each variable illustrated in Table 4. In this study, the action spaces of all agents are discrete. The action spaces of the pre-agents differ from those of the proto-agents, with the possible values of each agent at each dimension of action presented in Table 5. As can be inferred from Table 5, the pre-agents have only three possible values at each dimension of action, indicating that the action space of the pre-agents has significantly fewer possible values at each dimension of action compared to that of the proto-agents.

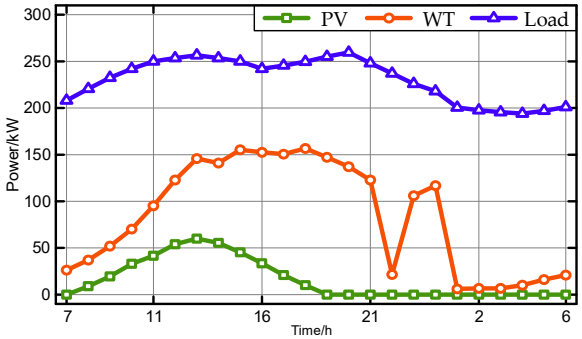


**Figure 7.** The PV power, WT power, and load data within the microgrid.

**Table 2.** Time-of-use pricing for electricity purchase and sale of the microgrid from the external grid.

| Time/h | Electricity Purchase Price (CNY/(kW·h)) | Electricity Sales Price (CNY/(kW·h)) |
|---|---|---|
| 1–6, 22–24 | 0.37 | 0.28 |
| 7–9, 14–17, 20, 21 | 0.82 | 0.65 |
| 10–13, 18, 19 | 1.36 | 0.78 |

**Table 3.** The relevant parameters of the devices in the microgrid.

| Main Parameters | Values | Main Parameters | Values |
|---|---|---|---|
| $P_{\mathrm{EV},\,i}^{\min}$ | $0.02 \cdot E_{\mathrm{EV},i}$ kW | $R_{1\mathrm{down}}$ | 80 kW |
| $P_{\mathrm{EV},\,i}^{\max}$ | $E_{\mathrm{EV},i}$ kW | $R_{1\mathrm{up}}$ | 80 kW |
| $P_{\mathrm{MT}}^{\min}$ | 20 kW | $E_{\mathrm{ES}}$ | 500 kW·h |
| $P_{\mathrm{MT}}^{\max}$ | 200 kW | $SOC_{\mathrm{ES}}(0)$ | 0.5 |
| $P_{\mathrm{ES}}^{ch.\max}$ | 50 kW | $\alpha$ | 0.0013 |
| $P_{\mathrm{ES}}^{dis.\max}$ | −50 kW | $\beta$ | 0.553 |
| $SOC_{\mathrm{EV},\,i}^{\min}$ | 0.1 | $c$ | 14.17 |
| $SOC_{\mathrm{EV},\,i}^{\max}$ | 1 | $g_E$ | 0.5 |
| $SOC_{\mathrm{ES}}^{\min}$ | 0.1 | $g_P$ | 10 |
| $SOC_{\mathrm{ES}}^{\max}$ | 0.9 | $\eta_{\mathrm{EV}}$, $\eta_{\mathrm{ES}}$ | 1 |
| $P_{\mathrm{nom}}$ | 60 kW | $\zeta_{\mathrm{EV}}$, $\zeta_{\mathrm{ES}}$ | 1 |

**Table 4.** The upper and lower boundaries of each variable in the state space.

| Variables | Agent | Lower Boundary | Upper Boundary |
|---|---|---|---|
| $t$ | EV and DG | 0:00 | 24:00 |
| $P_{\mathrm{EV,sum}}$ | EV | 0 kW | 750 kW |
| $P_{\mathrm{L}}$ | EV | 0 kW | 500 kW |
| $P_{\mathrm{grid}}$ | EV | −1000 kW | 1000 kW |
| $\sigma^b$ | EV | 0 CNY | 2 CNY |
| $\sigma^s$ | EV | 0 CNY | 2 CNY |
| $P_{\mathrm{PV}}$ | DG | 0 kW | 200 kW |
| $P_{\mathrm{WT}}$ | DG | 0 kW | 300 kW |
| $P_{\mathrm{MT}}$ | DG | 0 kW | 200 kW |
| $P_{\mathrm{ES}}$ | DG | −50 kW | 50 kW |
| $SOC_{\mathrm{ES}}$ | DG | 0 | 1 |

**Table 5.** All possible values for each variable in the action space.

| Variables | Agent | All Possible Values |
|---|---|---|
| $P_{ref}^{pre}$ | Pre-training EV | {3, 15, 30} |
| $SOC_{ref}^{pre}$ | Pre-training EV | {0.8, 0.9, 1} |
| $P_{\mathrm{MT},\,ref}^{pre}$ | Pre-training DG | {20, 110, 200} |
| $P_{\mathrm{ES},\,ref}^{pre}$ | Pre-training DG | {−50, 0, 50} |
| $P_{ref}$ | EV | {3, 6, 9, 12, ⋯, 60} |
| $SOC_{ref}$ | EV | {0.8, 0.82, 0.84, 0.86, ⋯, 1} |
| $P_{\mathrm{MT},\,ref}$ | DG | {20, 25, 30, 35, ⋯, 200} |
| $P_{\mathrm{ES},\,ref}$ | DG | {−50, −45, ⋯, −5, 0, 5, ⋯, 45, 50} |

We employ the pretraining method proposed in this study for 500 episodes. During the pretraining stage, the pre-agents interact with the environment to identify the optimal scheduling strategy of the microgrid under the pretrained action space. Proto-agents mimic the actions of their respective pre-agents. After the completion of pretraining, we extract the pretrained proto-agents and allow them to interact with the environment for 800 episodes to seek the optimal scheduling strategy of the microgrid under the prototype action space. The reward curves of the agent training process during the pretraining and training stages are illustrated in Figures 8 and 9, respectively.

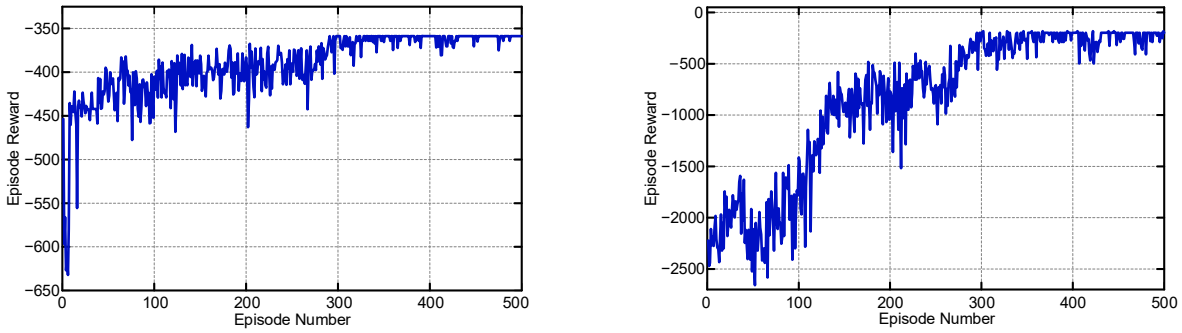


**Figure 8.** The reward curve of pre-agents and proto-agents during pretraining stage. (**a**) The reward curve of pre-agents during pretraining. (**b**) The reward curve of proto-agents during pretraining.

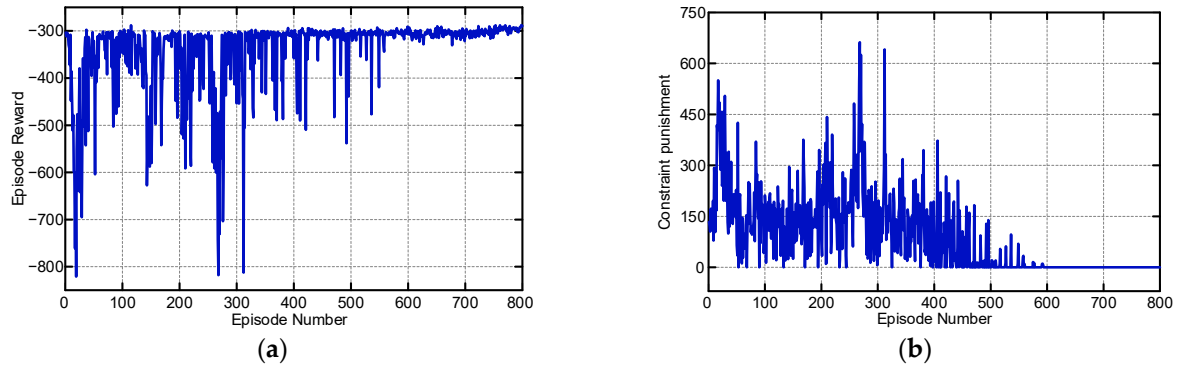


**Figure 9.** Reward curve and constraint punishment curve for agents in the training stage. (**a**) Reward curve. (**b**) Constraint punishment curve.

Figure 8 depicts the reward curves of the pre-agents and the proto-agents during the pretraining stage. As seen in Figure 8a, during the pretraining stage, due to the smaller action space of the pre-agents, they are able to quickly find the optimal operating strategy for the microgrid. Concurrently, as evidenced by Figure 8b, through the process of pretraining, the proto-agents are gradually able to reduce the disparity in output actions with the pre-agents, thereby initializing the neural networks of the proto-agents. After 300 episodes of pretraining, the reward curves of the proto- and the pretrained agents essentially stabilize. After 500 episodes of pretraining, we retain the proto-agents, allowing them to interact with the microgrid environment to compute the optimal operating strategy for the microgrid under prototype action space. As indicated by Figure 9, after 600 training episodes, the reward curve of the proto-agents essentially converges, signifying that the agents have discovered the optimal operating strategy for the microgrid under the prototype action space. After training the proto-agents for 800 episodes, we acquire the well-trained proto-agents. We deploy them for decision-making in the microgrid environment, obtaining the 24 h control reference signals provided by the EV and DG agents as illustrated in Figure 10.

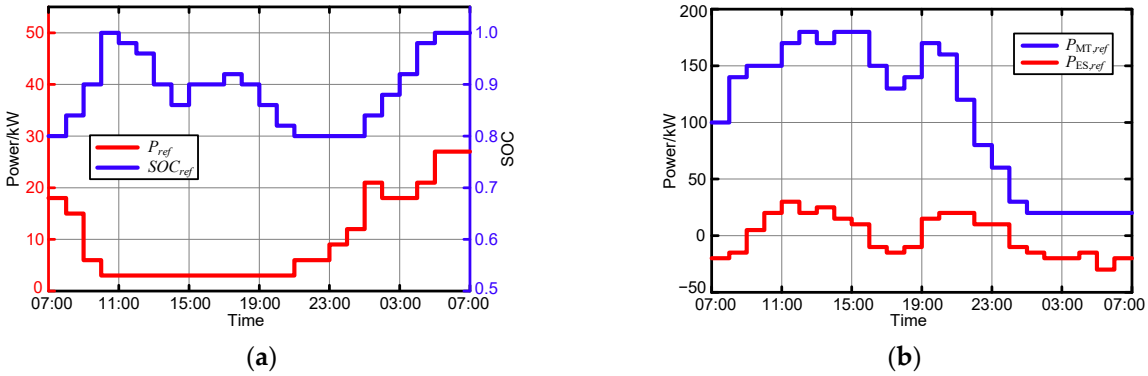


**Figure 10.** The 24 h control reference signals provided by the EV and DG agents. (**a**) Control reference signal provided by the EV agent. (**b**) Control reference signal provided by the DG agent.

### *4.1. Analysis of Optimized Operation Results for EVs*

Based on the control reference signals from the EV and DG agents, the charging and discharging processes of the 12 EVs during the time they are connected to the microgrid are depicted in Figure 11.

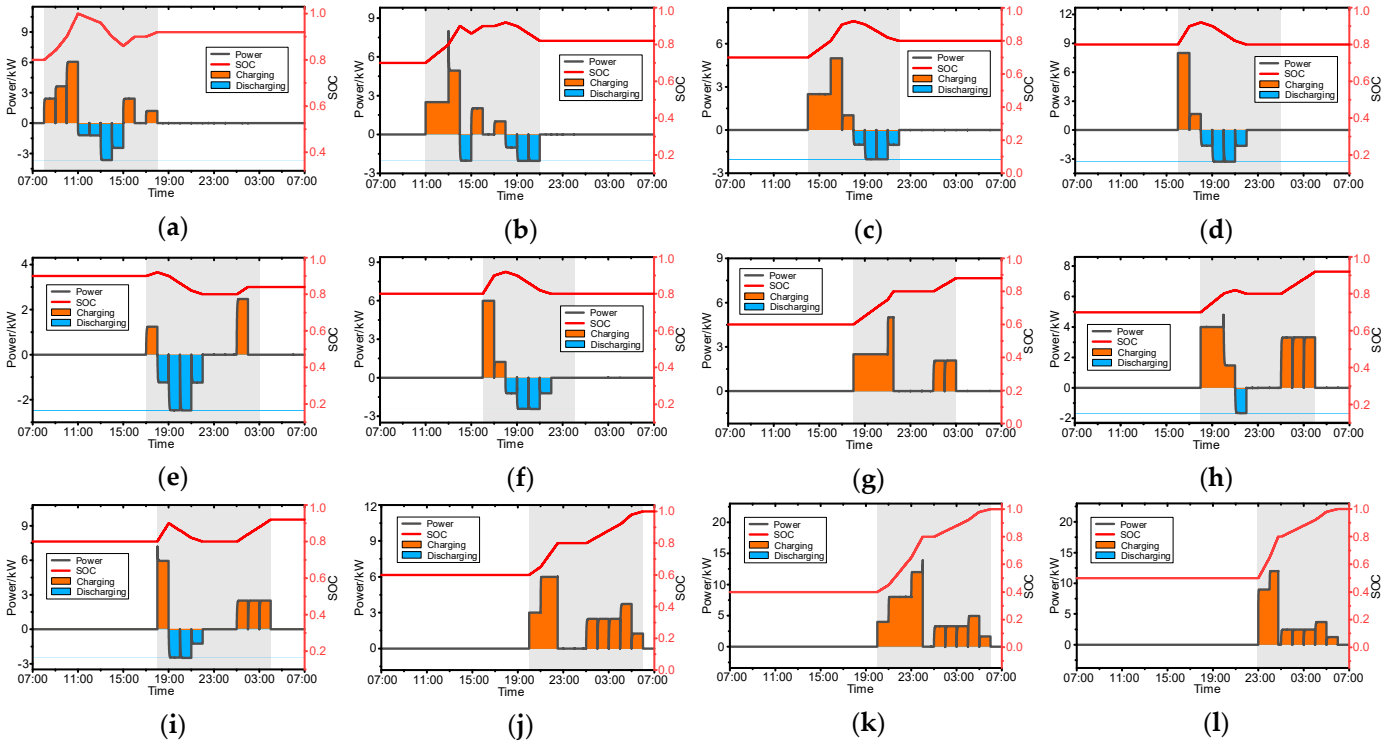


**Figure 11.** The charging and discharging processes of the 12 EVs during the time they are connected to the microgrid. (**a**) EV1. (**b**) EV2. (**c**) EV3. (**d**) EV4. (**e**) EV5. (**f**) EV6. (**g**) EV7. (**h**) EV8. (**i**) EV9. (**j**) EV10. (**k**) EV11. (**l**) EV12.

According to Table 6, the time when the SOC of 12 EVs is greater than or equal to 0.8 as a percentage of the total time connected to the microgrid can be obtained based on Figure 11. From Table 6, it can be inferred that the two-stage control architecture ensures that the SOC of all EVs exceeds 0.8 for over 60% of the time they are connected to the microgrid. In other words, except for a small period of time when EVs are first connected to the microgrid, the EVs can meet daily travel needs at any other time. Therefore, this two-stage control architecture fully respects the rights and interests of EV owners, enabling them to freely use their EVs for longer periods of time. In addition, the situation where

the SOC of EVs is less than 0.8 only exists for a period of time immediately after EVs are connected to the microgrid, and this charging method is also more in line with the driving habits of most vehicle owners. Finally, due to the stipulations within this paper, microgrids exclusively levy charges for the electrical energy consumed by EVs during the constrained scheduling phase, while abstaining from any fees for EV charging and discharging during the unrestricted scheduling phase. Consequently, Table 6 reveals that the billing duration for all EVs remains below 40% of the total time they are connected to the microgrid. Notably, for EVs 1, 4, 5, 6, and 9, the microgrid refrains from imposing any charges. Furthermore, during the constrained scheduling phase, the optimization methodology proposed in this study restricts EVs to charging only, thereby precluding any increase in EV charging costs. In summary, the method posited in this paper ensures both EV participation in microgrid optimization scheduling and user satisfaction.

**Table 6.** The SOC of 12 EVs is greater than or equal to 0.8 as a percentage of the total time connected to the microgrid.

| EV No. | Percentage | EV No. | Percentage |
|---|---|---|---|
| 1 | 100% | 7 | 60.97% |
| 2 | 80.19% | 8 | 79.97% |
| 3 | 76.21% | 9 | 100% |
| 4 | 100% | 10 | 74.90% |
| 5 | 100% | 11 | 63.31% |
| 6 | 100% | 12 | 77.96% |

To demonstrate the superiority of the method for optimizing the operation of EVs proposed in this article, we compare it with the traditional charging method for EVs in microgrids. In the traditional charging method for EVs in microgrids, all EVs start charging immediately upon arrival at the microgrid until they are fully charged. We assume that the charging speed of each EV is such that the SOC of the EV increases by 0.2 per hour. The load level of the microgrid with and without EVs within 24 h is shown in Figure 12. From Figure 12, it can be seen that the connection of EVs will greatly increase the load fluctuation of the microgrid and increase the operational risk of the microgrid. In addition, combined with Table 2, it can be seen that the load added by EVs charging is mostly concentrated in high electricity price periods, which indicates that the connection of EVs will also greatly increase the cost of purchasing electricity from external power grids for the microgrid. The load level of the microgrid obtained by the optimization operation method proposed in this paper is shown in Figure 13. By comparing it with the load level of a microgrid without an EV scheduling strategy for 24 h, we find that the scheduling strategy obtained in this article can reduce the load level from 12 to 21 and transfer the load during this period through EV charging and discharging to other periods. This reduces the load fluctuation of the microgrid and also reduces its electricity purchasing cost from external power grids.

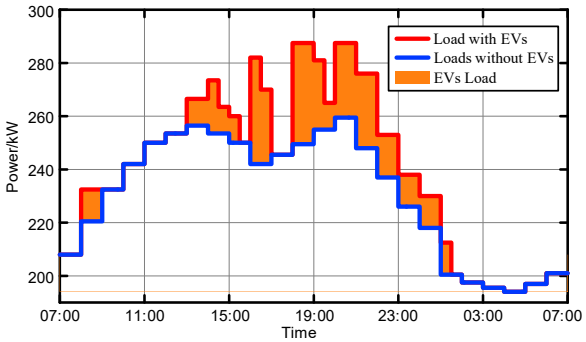


**Figure 12.** The load level of the microgrid with and without EVs within 24 h.

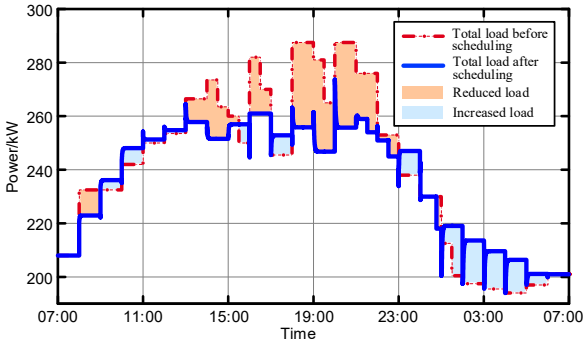


**Figure 13.** The load level of the microgrid obtained by the optimization operation method proposed in this paper.

### *4.2. Analysis of Lower-Layer Consistency Control Effectiveness in the Microgrid*

To validate the effectiveness of the lower-layer consensus control for EVs within the microgrid, we use as examples the charging and discharging power and SOC changes of EVs numbered 7, 8, and 9. As illustrated in Figure 14, these three EVs connect to the microgrid at 17:00, with their initial SOCs at the connection being 0.6, 0.7, and 0.8, respectively. The capacity ratios of the three EVs are 5:8:6. Upon connection, EVs 7 and 8 enter the constrained scheduling stage, as their SOC is less than 0.8. EV 9, whose SOC equals 0.8 at connection, enters the free scheduling stage postconnection. As depicted in Figure 14a, the output power of the two EVs in the constrained scheduling stage is allocated in a 5:8 ratio. As can be seen in Figure 14b, the EV in the free scheduling stage is able to follow the reference signals provided by the upper-layer deep reinforcement learning agents effectively.

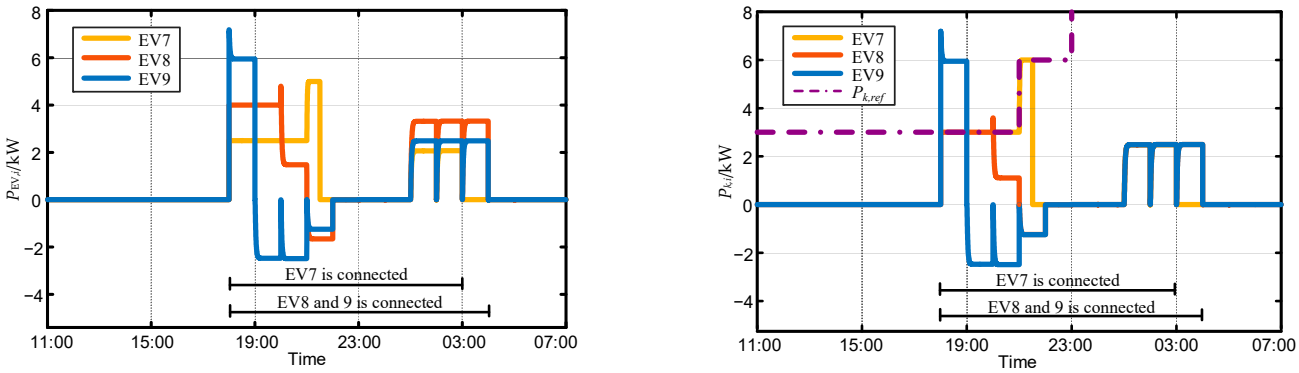


**Figure 14.** Power changes of the three EVs EV7, EV8, EV9. (**a**) Changes in $P_{\mathrm{EV},i}$ for three EVs. (**b**) Changes in $P_{k,i}$ for three EVs.

As depicted in Figure 15, the SOC of EV 7 reaches 0.8 between 21:00 and 22:00, and the SOC of EV 8 reaches 0.8 at 20:00. Once the SOC of the EVs reaches 0.8, they enter the free scheduling stage and autonomously adjust their output power according to the SOC reference signals. From Figure 15, it can be observed that during the free scheduling stage, the three EVs can meet the SOC reference signals at the end of each scheduling timestep, with the SOC linearly increasing or decreasing within each scheduling timestep. In addition, after entering the free scheduling stage, the SOC of all three EVs do not fall below 0.8.

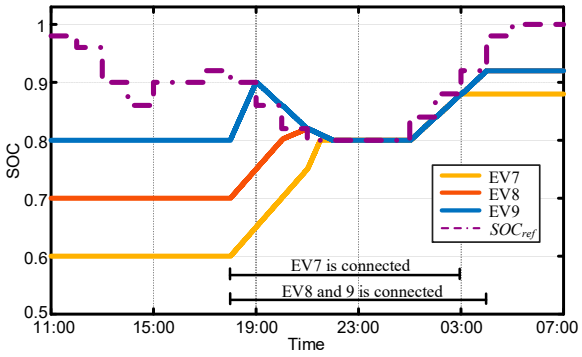


**Figure 15.** Changes in SOC of the three EVs EV7, EV8, and EV9.

In conclusion, based on the foregoing analysis, it can be inferred that the lower-layer control method for EVs proposed in this paper is capable of ensuring that the output power of EVs in the constrained scheduling stage aligns with the power reference signal $P_{ref}$ provided by the upper-layer EV agents. Concurrently, it allows the SOC of EVs in the free scheduling stage to undergo linear variations, achieving the SOC reference values provided by the EV agent at the end of each scheduling timestep.

### *4.3. Analysis of Optimized Scheduling Results in the Upper Layers of the Microgrid*

To analyze the economic and safety aspects of the scheduling strategies provided by the two-stage multiagent reinforcement learning method proposed in this paper, the trained EV and DG agents are deployed within the designed microgrid framework of this study. The power of various forms of energy and load conditions at each time period are depicted in Figure 16. The power of EVs in Figure 16 refers to the total power of all EVs. For convenience of representation, the average power in each time period is used here to represent the power of EVs and the grid during that period.

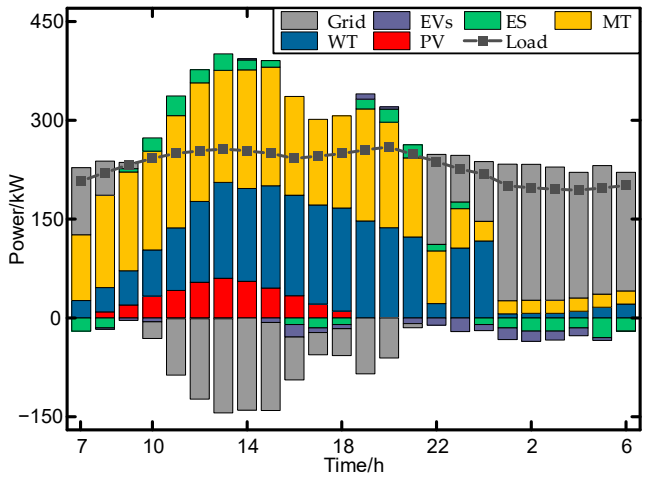


**Figure 16.** The power of various forms of energy and load conditions at each time period.

As is evident from Figure 16 and Table 2, the period at 7:00 is a high-electricity-price period, and 8:00–11:00 is a medium-electricity-price period. From 7:00 to 11:00, due to ramping constraints, the output power of the MT gradually increases from 100 kW. The external grid supplies power to the microgrid to meet the microgrid's load demand. Simultaneously, the EVs and ES charge during this period, utilizing the lower-priced electricity. From 11:00 to 15:00 is a high-electricity-price period, during which the MT's output power reaches its maximum, and the ES discharges while the microgrid sells power to the external grid for higher profits.

Between 16:00 and 19:00, another medium-electricity-price period, the output of the MT decreases, but the renewable energy output within the microgrid is substantial. During this period, the microgrid buys and sells less power from the external grid, relying more on internal MTs and renewable energy output to meet its load demand. Meanwhile, ES and EVs within the microgrid charge utilizing the internal electricity. From 19:00 to 20:00, it reenters a high-electricity-price period, the output of the MT increases, EVs and ES discharge, and the microgrid sells power to the external grid for profit. From 21:00 to 22:00, a medium-electricity-price period, the MT output decreases, but due to the high load level and a considerable number of EVs in the microgrid, the EVs and ES discharge to reduce the internal load demand. From 23:00 to 6:00 the next day, it enters a low-electricity-price period again, where the MT operates at its lowest output power. The microgrid uses low-cost electricity purchased from the external grid to power its internal loads and charge the ES device and EVs.

The output power of the MT in the microgrid is shown in Figure 17. As seen in Figure 17, due to the impact of the constraint penalty Formula (31), the output power of the MT at two adjacent moments does not exceed the MT's ramp-up and ramp-down constraints. This implies that the scheduling strategy provided by the method proposed in this paper can ensure the safe operation of the MT. The charge/discharge power and SOC change conditions of the ES are illustrated in Figure 18. From Figure 18, it is evident that throughout the entire scheduling cycle, the SOC of the ES device has neither exceeded nor fallen below its limits. Moreover, due to the impact of the constraint penalty Formula (33), at the end of the scheduling cycle, the SOC value of the ES is able to return to its initial value at the beginning of the scheduling. This demonstrates that the scheduling strategy provided by the method proposed in this paper can guarantee the safe operation of the ES device.

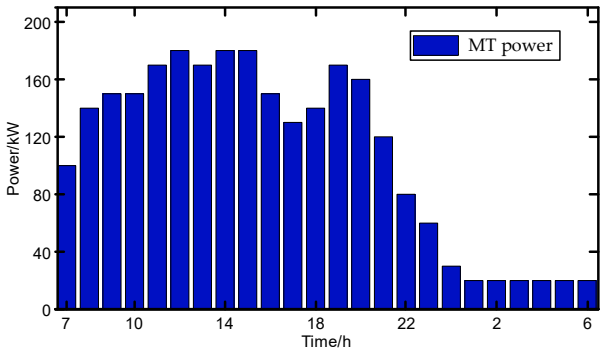


**Figure 17.** The output power of the MT in the microgrid.

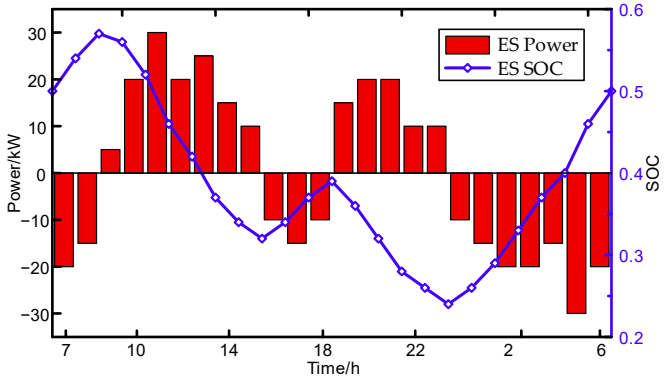


**Figure 18.** The charge/discharge power and SOC change conditions of the ES.

Through the aforementioned analysis, it can be observed that the scheduling strategy obtained through the two-stage PPO method proposed in this paper is capable of reasonably

coordinating the MT, ES, and EVs within the microgrid. It ensures the secure operation of the microgrid while optimizing economic benefits significantly.

### *4.4. Comparative Analysis*

To validate the enhancements achieved by the PPO pretraining method proposed in this paper over the original algorithm, this section contrasts the training processes of the MAPPO method augmented with our pretraining approach against the traditional MAPPO method. The reward curves for both training procedures are illustrated in Figure 19. As observed from Figure 19, the MAPPO algorithm, without the incorporation of the proposed pretraining approach, fails to identify the optimal scheduling strategy within 800 episodes, and it exhibits substantial fluctuations in its reward curve during training. Additionally, between episodes 523 and 773, the agent's output actions become entrapped in a local optimum. By introducing the pretraining method proposed in this paper, there is a noticeable acceleration in the convergence speed of the reward curve, which also exhibits reduced oscillations, making the training process significantly more stable. From this, we can conclude that the pretraining approach presented in this paper enhances the algorithm's rate of convergence, elevates the stability during the training stage, and prevents the agent's output actions from converging to local optima.

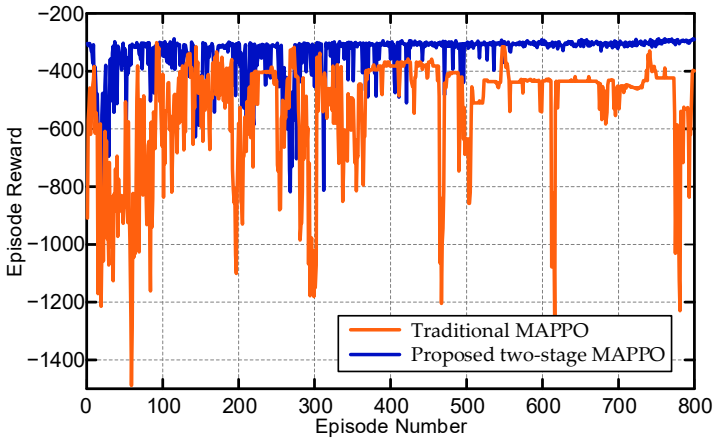


**Figure 19.** Comparison of training process reward curves for traditional MAPPO and the two-stage MAPPO proposed in this paper.

The two-stage MAPPO method proposed in this paper is compared to other deep reinforcement learning algorithms applied to the microgrid scheduling problem in discrete spaces, including the PPO algorithm [35], DQN algorithm [36], dueling deep Q-network (DDQN) algorithm [37], dueling double deep Q-network (D3QN) algorithm [38], advantage actor–critic (A2C) algorithm [39], and asynchronous advantage actor–critic (A3C) algorithm [40]. These algorithms are each applied to the microgrid model designed in this paper for training for 800 episodes. The reward curves for each of these deep reinforcement learning algorithms are depicted in Figure 20, and the average rewards of the last 20 episodes of the training process for each algorithm are illustrated in Figure 21. From Figures 20 and 21, it is evident that the two-stage MAPPO method proposed in this study is capable of deriving the optimal scheduling strategies compared to other methods. Concurrently, the proposed method manifests a faster convergence speed and smaller fluctuations in the reward curves during the training process. This indicates that the two-stage MAPPO method proposed in this paper exhibits superior optimization performance in addressing microgrid scheduling problems.

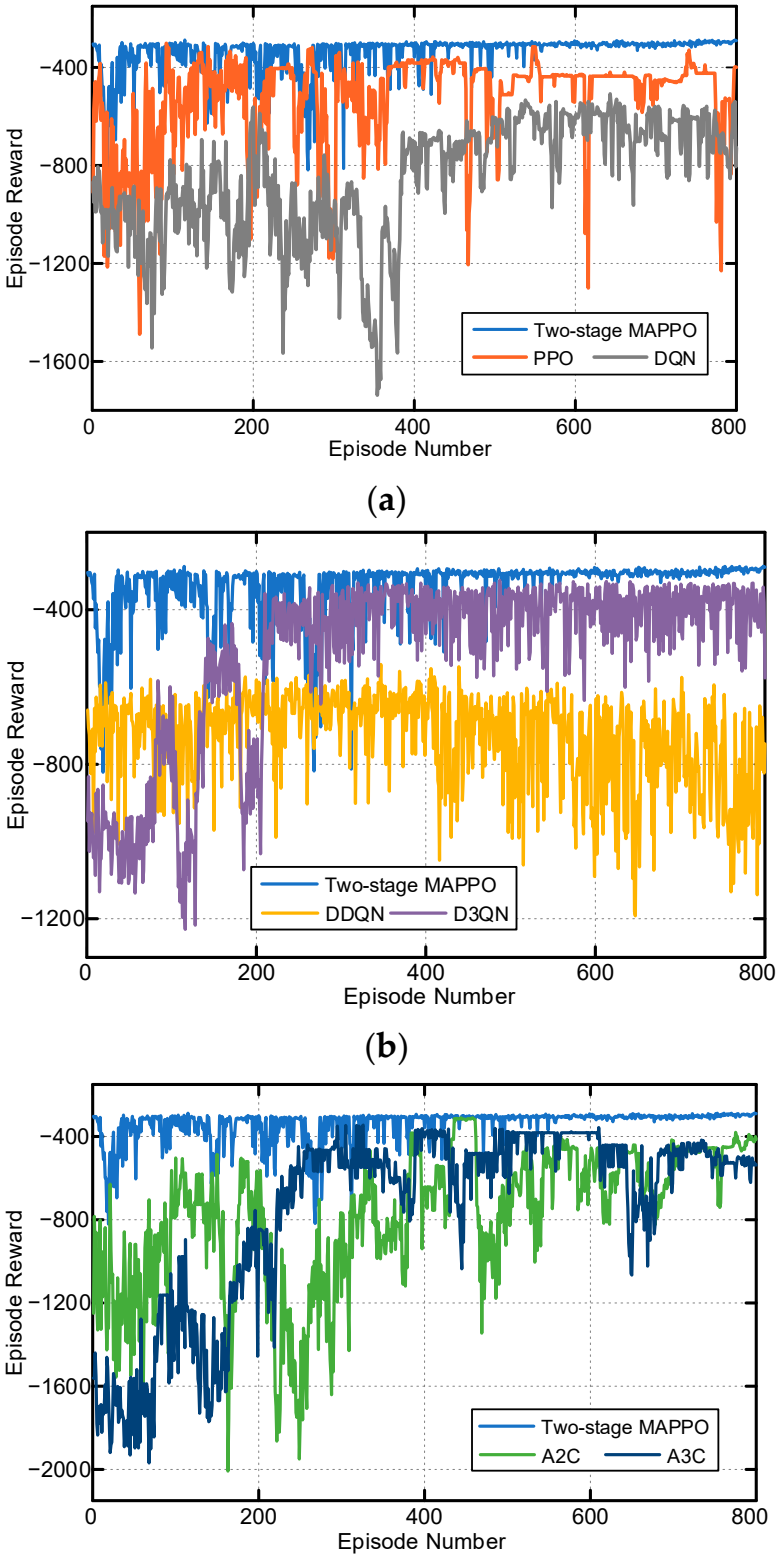


**Figure 20.** Comparison of reward curves of different deep reinforcement learning algorithms applied to the microgrid environment designed in this paper. (**a**) The proposed method is compared with PPO and DQN algorithms. (**b**) The proposed method is compared with DDQN and D3QN algorithms. (**c**) The proposed method is compared with A2C and A3C algorithms.

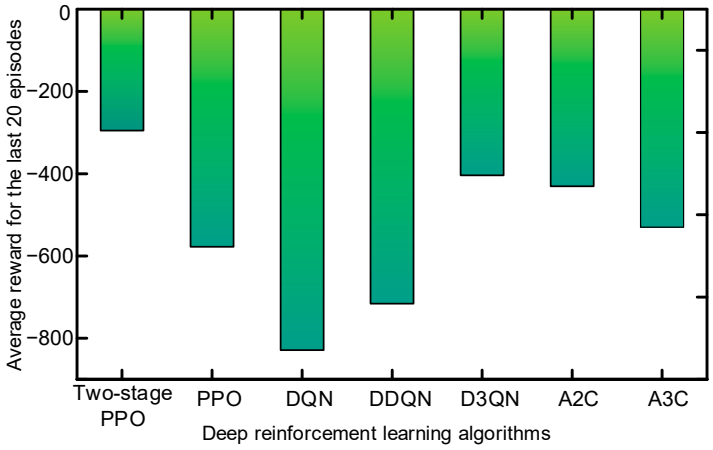


**Figure 21.** The average rewards of the last 20 episodes of the training process for each deep reinforcement learning algorithm.

### *4.5. Robustness Analysis*

When optimizing scheduling methods for practical applications, environmental parameters of microgrids and the operating status of each EV are often difficult to predict. Therefore, fixed microgrid operational strategies calculated based on predicted data often fail to achieve the expected results in practical applications. This requires scheduling algorithms to have higher robustness, which means that the algorithm can autonomously adjust its own action strategy according to different environmental conditions. Compared with traditional methods used for microgrid optimization decisions (evolutionary algorithms, swarm intelligence algorithms), reinforcement learning methods often demonstrate stronger robustness. This is because during the training process of the reinforcement learning agent, the agent will continue to try to accumulate experience through its internal neural network, which will become a useful resource for the agent in actual environments, helping the agent better understand and respond to environmental changes. This allows the reinforcement learning agent to adaptively adjust its own output action strategy based on the experience gained during the training process when the environment changes, thereby exhibiting stronger robustness. Since the two-stage MAPPO algorithm proposed in this paper enables the intelligent agent to find the optimal operating strategy of the microgrid faster and more stably, it further improves the robustness of the reinforcement learning agent to a certain extent. To verify the advantage of the two-stage MAPPO method proposed in this paper in terms of operational decision robustness, we designed four different microgrid operating scenarios in this section:

- Scenario 1: The microgrid scenario provided above.
- Scenario 2: EVs 3, 5, 7, and 10 arrive one hour early.
- Scenario 3: The initial SOC of EVs 3, 5, 7, and 10 is 0.1 less than the data provided above.
- Scenario 4: EVs 3, 5, 7, and 10 arrive one hour early and their initial SOC is 0.1 less than the data provided above.

We compared our two-stage MAPPO method with the non-dominated sorting genetic algorithm III (NSGA-III) algorithm [41] and PSO algorithm [42], which are commonly used for microgrid optimization scheduling. It is worth noting that NSGA-III algorithm is typical of evolutionary algorithms and the PSO algorithm is typical of swarm intelligence algorithms. We applied these three methods to four different microgrid scenarios and observed the rewards obtained by the microgrid environment model. The results are shown in Figure 22.

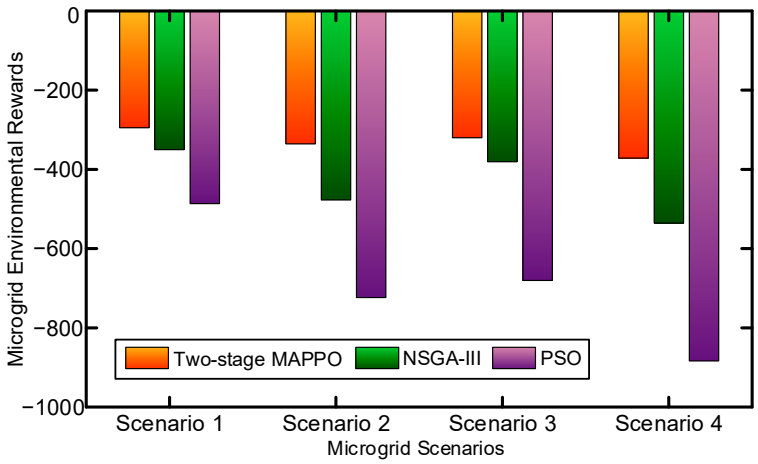


**Figure 22.** Comparison of the rewards of the three algorithms applied to different microgrid scenarios.

As shown in Figure 22, compared with the NSGA-III algorithm and PSO algorithm, our proposed two-stage MAPPO algorithm achieved the highest rewards in all four different scenarios. This indicates that our proposed method can make relatively good action decisions when dealing with different microgrid scenarios. In other words, our proposed two-stage MAPPO method has stronger robustness.

## 5. Discussion

According to the simulation analysis in Section 4, we can conclude that the proposed two-stage, dual-layer optimization operation method for microgrids containing EVs has several advantages, as follows:

1. The proposed two-stage control architecture fully respects the rights and interests of EV owners by ensuring that if an EV is connected to the microgrid and its SOC is less than 0.8, it can only be charged until its SOC is greater than 0.8. Once the SOC is greater than 0.8, as analyzed in Section 3, the EV can support the car owner's daily travel needs. After the SOC is greater than 0.8, the proposed control method ensures that the EV's SOC will not drop below 0.8. In other words, before the EV can support the car owner's daily travel needs, it can only be charged. Once the EV's SOC rises to a level that can meet the car owner's daily travel needs, the car owner can use their EV at any time to meet their travel needs.
2. The proposed two-stage control architecture retains the ability of EVs to participate in microgrid optimization and scheduling. When an EV's SOC is less than 0.8, it participates in microgrid scheduling by adjusting its charging power; when its SOC is greater than 0.8, it can participate in microgrid scheduling by charging or discharging. Therefore, the proposed two-stage control architecture fully respects the charging rights of EV owners while retaining their ability to participate in microgrid optimization and scheduling.
3. The lower-level control ensures that all EVs are charged uniformly, avoiding conflicts caused by uneven charging of EVs.
4. The new two-stage MAPPO algorithm-based optimization scheduling method proposed in the upper level has faster calculation speed and better calculation effect than traditional single-stage MAPPO algorithms because we added a new pretraining stage for the traditional single-stage MAPPO algorithm. In this pretraining stage, we introduced pre-agents which have fewer action choices in each dimension of action space, meaning they can find optimal scheduling strategies more quickly. Therefore, using proto-agents' neural network trained with output actions from pre-agents as a pretraining stage significantly improves calculation speed and effect during agent training compared to traditional single-stage MAPPO algorithms.

5. The proposed dual-layer optimization operation method reduces microgrid operating costs while ensuring the safe operation of microgrids containing EVs.

Attention should be paid to the fact that during the constraint scheduling phase, the methods proposed may, in order to ensure the overall benefits of the microgrid, reduce the charging power of EVs significantly, thereby extending the charging time. This trade-off may potentially diminish the satisfaction of EV owners regarding charging. However, it represents a balanced choice in this article, considering both the interests of EV owners and the overall benefits of the microgrid. The fact that scheduling plans still inadvertently prolong charging durations necessitates the formulation of a more judicious compensation mechanism for EV users. Such a mechanism should holistically consider factors including charging tariffs, initiation time for charging, duration of microgrid connection, and construction costs of microgrid charging infrastructure. Given the primary focus of this discourse on microgrid operational costs and scheduling directives, we shall refrain from excessive elaboration on the EV user compensation mechanism. In subsequent research endeavors, we intend to delve further into the compensation framework for EV participation in microgrids. However, it remains undeniable that, as articulated in Section 4.1, the method proposed herein has significantly curtailed EV billing durations compared to conventional approaches, thereby ensuring a certain degree of satisfaction among EV users. In summary, our proposed two-stage, dual-layer optimization operation method considers both the rights and interests of EV owners and the overall benefits of microgrids containing EVs, providing a new solution to optimize operation problems for urban microgrids containing EVs.

## 6. Conclusions

This paper proposes a two-stage, dual-layer optimization operational method tailored for microgrids incorporating EVs. The method combines consistency control methods with deep reinforcement learning methods and simultaneously considers the overall benefits of microgrid optimization scheduling with EV participation and the charging benefits of EV owners. In summary, the main innovations of this paper are as follows:

1. A two-stage control architecture is designed to solve the conflict between EV participation in microgrid optimization scheduling and respecting the charging benefits of EV owners.
2. A two-layer optimization operation method is proposed, which quickly and accurately finds the optimal operating strategy of the microgrid while ensuring that all EVs operate under consistent conditions.
3. Taking into account the different charging objectives of EVs in different charging stages, a new two-stage consistency control method suitable for the two-stage EV control architecture is proposed based on the traditional consistency method. This helps to ensure EV consistent power charging in the constrained scheduling stage and consistent SOC changes in the free scheduling stage.
4. A new two-stage MAPPO algorithm is proposed by leveraging the speed and accuracy of reinforcement learning optimization in a discrete action space. The introduction of a pretrained agent in the discrete action space guides the agent to make optimal decisions in the discrete action space during the pretraining phase and initializes the neural network of the agent.

We hope that this study will serve as a valuable reference and guide for the construction of future urban microgrids. In future work, we will conduct more in-depth research on optimizing operation problems for microgrids containing larger-scale and different types of EVs.

**Author Contributions:** Conceptualization, B.Z. and Z.Z.; methodology, B.Z. and Z.Z.; software, Z.Z. and C.X.; validation, Z.Z. and C.X.; formal analysis, B.Z. and B.L.; data curation, B.Z. and Z.Z.; writing—original draft preparation, B.Z. and C.X.; writing—review and editing, B.Z. and Z.Z.; visualization, Z.Z. and B.L.; supervision, B.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in part by the National Natural Science Foundation of China, grant number U22B20115, in part by the Applied Fundamental Research Program of Liaoning Province, grant number 2023JH2/101600036.

**Data Availability Statement:** The supporting data for this study can be found within this paper.

**Acknowledgments:** Special thanks to the Intelligent Electrical Science and Technology Research Institute, Northeastern University (China), for providing technical support for this research.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

Full terms corresponding to acronyms mentioned in this paper:

| | |
|---|---|
| EV | Electric vehicle |
| DG | Distributed generation |
| SOC | State of charge |
| MT | Microturbine |
| ES | Energy storage |
| PV | Photovoltaic |
| WT | Wind turbine |
| MDP | Markov decision process |
| EU | European Union |
| HOV | High-occupancy vehicle |
| NSGA-II | Non-dominated sorting genetic algorithm II |
| NSGA-III | Non-dominated sorting genetic algorithm III |
| PSO | Particle swarm optimization |
| DQN | Deep Q-network |
| DDPG | Deep deterministic policy gradient |
| MAPPO | Multiagent proximal policy optimization |
| DDQN | Dueling deep Q-network |
| D3QN | Dueling double deep Q-network |
| A2C | Advantage actor–critic |
| A3C | Asynchronous advantage actor–critic |

## References

1. Ferro, G.; Laureri, F.; Minciardi, R.; Robba, M. A predictive discrete event approach for the optimal charging of electric vehicles in microgrids. *Control Eng. Pract.* **2019**, *86*, 11–23. [CrossRef]
2. Guo, S.; Li, P.; Ma, K.; Yang, B.; Yang, J. Robust energy management for industrial microgrid considering charging and discharging pressure of electric vehicles. *Appl. Energy* **2022**, *325*, 119846. [CrossRef]
3. Khan, I.S.; Ahmad, M.O.; Majava, J. Industry 4.0 and sustainable development: A systematic mapping of triple bottom line, Circular Economy and Sustainable Business Models perspectives. *J. Clean. Prod.* **2021**, *297*, 126655. [CrossRef]
4. Sun, X. Green city and regional environmental economic evaluation based on entropy method and GIS. *Environ. Technol. Innov.* **2021**, *23*, 101667. [CrossRef]
5. Li, S.; Zhao, P.; Gu, C.; Li, J.; Cheng, S.; Xu, M. Battery protective electric vehicle charging management in renewable energy system. *IEEE Trans. Ind. Inf.* **2022**, *19*, 1312–1321. [CrossRef]
6. Bistline, J.E.; Young, D.T. The role of natural gas in reaching net-zero emissions in the electric sector. *Nat. Commun.* **2022**, *13*, 4743. [CrossRef]
7. Szumska, E.M. Electric Vehicle Charging Infrastructure along Highways in the EU. *Energies* **2023**, *16*, 895. [CrossRef]

8. Das, P.K.; Bhat, M.Y. Global electric vehicle adoption: Implementation and policy implications for India. *Environ. Sci. Pollut. Res.* **2022**, *29*, 40612–40622. [CrossRef]
9. Maghfiroh, M.F.N.; Pandyaswargo, A.H.; Onoda, H. Current readiness status of electric vehicles in indonesia: Multistakeholder perceptions. *Sustainability* **2021**, *13*, 13177. [CrossRef]
10. Huang, Y.; Masrur, H.; Lipu, M.S.H.; Howlader, H.O.R.; Gamil, M.M.; Nakadomari, A.; Mandal, P.; Senjyu, T. Multi-objective optimization of campus microgrid system considering electric vehicle charging load integrated to power grid. *Sustain. Cities Soc.* **2023**, *98*, 104778. [CrossRef]
11. Zhou, W.; Xu, C.; Yang, D.; Peng, F.; Guo, X.; Wang, S. Research on Demand Response Strategy of Electric Vehicles Considering Dynamic Adjustment of Willingness Under P2P Energy Sharing. *Proc. CSEE* **2022**, 1–14. Available online: https://kns.cnki.net/kcms/detail/11.2107.TM.20221019.1907.005.html (accessed on 2 November 2023).
12. Zhang, L.; Sun, C.; Cai, G.; Huang, N.; Lv, L. Two-stage Optimization Strategy for Coordinated Charging and Discharging of EVs Based on PSO Algorithm. *Proc. CSEE* **2022**, *42*, 1837–1852.
13. Matrone, S.; Ogliari, E.; Nespoli, A.; Gruosso, G.; Gandelli, A. Electric Vehicles charging sessions classification technique for optimized battery charge based on machine learning. *IEEE Access* **2023**, *11*, 52444–52451. [CrossRef]
14. Mei, Z.; Zhao, T.; Xie, X. Hierarchical Fuzzy Regression Tree: A New Gradient Boosting Approach to Design a TSK Fuzzy Model. *Inf. Sci.* **2023**, *652*, 119740. [CrossRef]
15. Yu, R.; Chen, Y.H.; Han, B.; Zhao, H. A hierarchical control design framework for fuzzy mechanical systems with high-order uncertainty bound. *IEEE Trans. Fuzzy Syst.* **2020**, *29*, 820–832. [CrossRef]
16. Cai, G.; Jiang, Y.; Huang, N.; Yang, D.; Pan, X.; Shang, W. Large-scale Electric Vehicles Charging and Discharging Optimization Scheduling Based on Multi-agent Two-level Game Under Electricity Demand Response Mechanism. *Proc. CSEE* **2023**, *43*, 85–99.
17. Arias, N.B.; Sabillón, C.; Franco, J.F.; Quirós-Tortós, J.; Rider, M.J. Hierarchical optimization for user-satisfaction-driven electric vehicles charging coordination in integrated MV/LV networks. *IEEE Syst. J.* **2022**, *17*, 1247–1258. [CrossRef]
18. Yao, C.; Chen, S.; Yang, Z. Joint routing and charging problem of multiple electric vehicles: A fast optimization algorithm. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8184–8193. [CrossRef]
19. Liu, P.; Wang, C.; Hu, J.; Fu, T.; Cheng, N.; Zhang, N.; Shen, X. Joint route selection and charging discharging scheduling of EVs in V2G energy network. *IEEE Trans. Veh. Technol.* **2020**, *69*, 10630–10641. [CrossRef]
20. Tang, Q.; Xie, M.; Yang, K.; Luo, Y.; Zhou, D.; Song, Y. A decision function based smart charging and discharging strategy for electric vehicle in smart grid. *Mob. Netw. Appl.* **2019**, *24*, 1722–1731. [CrossRef]
21. Qureshi, U.; Ghosh, A.; Panigrahi, B.K. Real-Time Control for Charging Discharging of Electric Vehicles in a Charging Station with Renewable Generation and Battery Storage. In Proceedings of the 2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET), Hyderabad, India, 21 January 2021.
22. Wang, Y.; Dong, W.; Yang, Q. Multi-stage optimal energy management of multi-energy microgrid in deregulated electricity markets. *Appl. Energy* **2022**, *310*, 118528. [CrossRef]
23. Yang, M.; Zhang, L.; Dong, W. Economic benefit analysis of charging models based on differential electric vehicle charging infrastructure subsidy policy in China. *Sustain. Cities Soc.* **2020**, *59*, 102206. [CrossRef]
24. Darabi, Z.; Ferdowsi, M. Impact of plug-in hybrid electric vehicles on electricity demand profile. *Smart Power Grids* **2011**, *2012*, 319–349. [CrossRef]
25. Raugei, M.; Hutchinson, A.; Morrey, D. Can electric vehicles significantly reduce our dependence on non-renewable energy? Scenarios of compact vehicles in the UK as a case in point. *J. Clean. Prod.* **2018**, *201*, 1043–1051. [CrossRef]
26. Olfati-Saber, R.; Fax, J.A.; Murray, R.M. Consensus and cooperation in networked multi-agent systems. *Proc. IEEE* **2007**, *95*, 215–233. [CrossRef]
27. Liu, W.; Gu, W.; Sheng, W.; Meng, X.; Wu, Z.; Chen, W. Decentralized multi-agent system-based cooperative frequency control for autonomous microgrids with communication constraints. *IEEE Trans. Sustain. Energy* **2014**, *5*, 446–456. [CrossRef]
28. Gu, W.; Liu, W.; Zhu, J.; Zhao, B.; Wu, Z.; Luo, Z.; Yu, J. Adaptive decentralized under-frequency load shedding for islanded smart distribution networks. *IEEE Trans. Sustain. Energy* **2014**, *5*, 886–895. [CrossRef]
29. Woo, J.; Yu, C.; Kim, N. Deep reinforcement learning-based controller for path following of an unmanned surface vehicle. *Ocean Eng.* **2019**, *183*, 155–166. [CrossRef]
30. Dossa, R.J.; Huang, S.; Ontañón, S.; Matsubara, T. An empirical investigation of early stopping optimizations in proximal policy optimization. *IEEE Access* **2021**, *9*, 117981–117992. [CrossRef]
31. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arxiv* **2017**, arXiv:1707.06347.
32. Rui, S.U.N.; Hai-tao, Y.U.; Yong, D.U. Time space distribution characteristics of taxi shift in Beijing. *J. Transp. Syst. Eng. Inf. Technol.* **2014**, *14*, 221.
33. Guo, D.; Tang, L.; Zhang, X.; Liang, Y.C. Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2020**, *69*, 13124–13138. [CrossRef]
34. Zhao, P.; Wu, J.; Wang, Y.; Zhang, H. Operation optimization strategy of microgrid based on deep reinforcement learning. *Electr. Power Autom. Equip.* **2022**, *42*, 9–16.
35. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. *Energy* **2022**, *238*, 121873. [CrossRef]

36. Duan, J.; Yi, Z.; Shi, D.; Lin, C.; Lu, X.; Wang, Z. Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC–DC microgrids. *IEEE Trans. Ind. Inf.* **2019**, *15*, 5355–5364. [CrossRef]
37. Wang, C.; Mei, S.; Yu, H.; Cheng, S.; Du, L.; Yang, P. Unintentional islanding transition control strategy for three-/single-phase multimicrogrids based on artificial emotional reinforcement learning. *IEEE Syst. J.* **2021**, *15*, 5464–5475. [CrossRef]
38. Li, Y.; Wang, Z.; Xu, W.; Gao, W.; Xu, Y.; Xiao, F. Modeling and energy dynamic control for a ZEH via hybrid model-based deep reinforcement learning. *Energy* **2023**, *277*, 127627. [CrossRef]
39. Chen, D.; Chen, K.; Li, Z.; Chu, T.; Yao, R.; Qiu, F.; Lin, K. Powernet: Multi-agent deep reinforcement learning for scalable powergrid control. *IEEE Trans. Power Syst.* **2021**, *37*, 1007–1017. [CrossRef]
40. Shi, M.; Huang, Y.; Lin, H. Research on power to hydrogen optimization and profit distribution of microgrid cluster considering shared hydrogen storage. *Energy* **2023**, *264*, 126113. [CrossRef]
41. Zhou, B.; Liu, B.; Yang, D.; Cao, J.; Littler, T. Multi-objective optimal operation of coastal hydro-electrical energy system with seawater reverse osmosis desalination based on constrained NSGA-III. *Energy Convers. Manag.* **2020**, *207*, 112533. [CrossRef]
42. Zishan, F.; Akbari, E.; Montoya, O.D.; Giral-Ramírez, D.A.; Molina-Cabrera, A. Efficient PID Control Design for Frequency Regulation in an Independent Microgrid Based on the Hybrid PSO-GSA Algorithm. *Electronics* **2022**, *11*, 3886. [CrossRef]