energies

MDPI

Review

Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control

Qingyan Li [1], Tao Lin [1,*], Qianyi Yu [2], Hui Du [1], Jun Li [1] and Xiyue Fu [1]

[1] Hubei Engineering and Technology Research Center for AC/DC Intelligent Distribution Network, School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China
[2] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
* Correspondence: tlin@whu.edu.com

Abstract: With the ongoing transformation of electricity generation from large thermal power plants to smaller renewable energy sources (RESs), such as wind and solar, modern renewable power systems need to address the new challenge of the increasing uncertainty and complexity caused by the deployment of electricity generation from RESs and the integration of flexible loads and new technologies. At present, a high volume of available data is provided by smart grid technologies, energy management systems (EMSs), and wide-area measurement systems (WAMSs), bringing more opportunities for data-driven methods. Deep reinforcement learning (DRL), as one of the state-of-the-art data-driven methods, is applied to learn optimal or near-optimal control policy by formulating the power system as a Markov decision process (MDP). This paper reviews the recent DRL algorithms and the existing work of operational control or emergency control based on DRL algorithms for modern renewable power systems and control-related problems for small signal stability. The fundamentals of DRL and several commonly used DRL algorithms are briefly introduced. Current issues and expected future directions are discussed.

Keywords: data-driven; artificial intelligence; deep reinforcement learning; control; modern renewable power system

Citation: Li, Q.; Lin, T.; Yu, Q.; Du, H.; Li, J.; Fu, X. Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control. *Energies* **2023**, *16*, 4143. https://doi.org/10.3390/en16104143

Academic Editor: Ahmed Abu-Siada

Received: 13 March 2023
Revised: 5 May 2023
Accepted: 12 May 2023
Published: 17 May 2023

1. Introduction

An electric power system is a comprehensive system that includes energy generation, transmission, and transformation, as well as consumption of electricity and other components. It is crucial to the continuing smooth operation of modern society. Power system failures, such as prevalent power outages, will inevitably lead to substantial economic losses [1] and social instability. For instance, the power system outage that happened in the United States and Canada in 2003 caused an estimated USD 10 billion [2]. Therefore, it is crucial to maintain stability and ensure the reliability of power systems.

In recent decades, the power system has been experiencing ongoing transformation and reconstruction to be more intelligent, sustainable, and distributed. Power systems have been evolving toward the objective of depending on a greater proportion of high-efficiency renewable energy sources, such as wind and solar power, which brings growing complexity and uncertainty for both the generation and demand sides in operating and investment decision-making processes [3]. Power electronic converters are commonly used to connect the renewable energy source (RES) generators to the power grid. Due to the fact that RESs, especially the grid-following RESs, usually do not have the capability to actively respond to frequency changes, the system inertia decreases and frequency stability issues become increasingly prominent; therefore, as a consequence, frequency regulation (FR) becomes more complex. Currently, renewable generation covers all levels of the power system, including transmission, distribution, and micro-grids (MGs). The complexity of power networks is also raised by the addition of new types of electrical loads to the system, such as the fast rise of electric cars (EVs).

The deployment of advanced communication infrastructures and newly installed devices, such as phasor measurement units (PMUs) in transmission systems, advanced smart meters in distribution networks in power systems, and high-performance computational capabilities provide a solid basis for advanced control techniques. For instance, reinforcement learning (RL) is a model-free technique for updating control actions based on observing the operational conditions of power systems. The goal of RL is to interact with the environment to develop the optimal policy that will yield the maximum reward. Since the early 20th century, the most common RL algorithm, including Q-learning and state-action-reward-state-action (SARSA), have been used to control power systems.

The combination of RL and deep learning is known as deep reinforcement learning (DRL). In order to handle problems involving sequential decision-making, DRL approximates a function using a deep neural network (DNN). DRL has become a recognized approach in many fields, such as gaming [4–6], computer version (CV) [7], smart transportation [8], automatic piloting [9], and other fields with great success. In addition, DRL became active in the control field of power systems in recent years by describing the power system as a Markov decision process (MDP). For safety concerns, DRL is usually used in power grid simulators and updates the control actions to find a (near-)optimal control policy.

This paper reviews the recent DRL-based control application in a normal operating state, emergency state, and control-related applications. The main advantages of this research following a recent review of DRL for applications in power systems [10–12] are as follows:

- This paper focuses on the state-of-the-art DRL-based methods for operational control and emergency control in modern renewable power systems;
- Rather than presenting comprehensive applications for power systems, this paper focuses on the control field according to the current operating states and levels;
- The significant limitations and potential remedies of DRL-based approaches in power system control applications are thoroughly concluded and discussed.

The structure of this paper is as follows. In Section 2, the foundations of (D)RL and various widely used algorithms are presented. In Section 3, an overview of DRL-based methods for power system control applications is given. Section 4 presents the discussion and future directions, and Section 5 concludes with a summary. The structure and main content of this paper are shown in Figure 1.

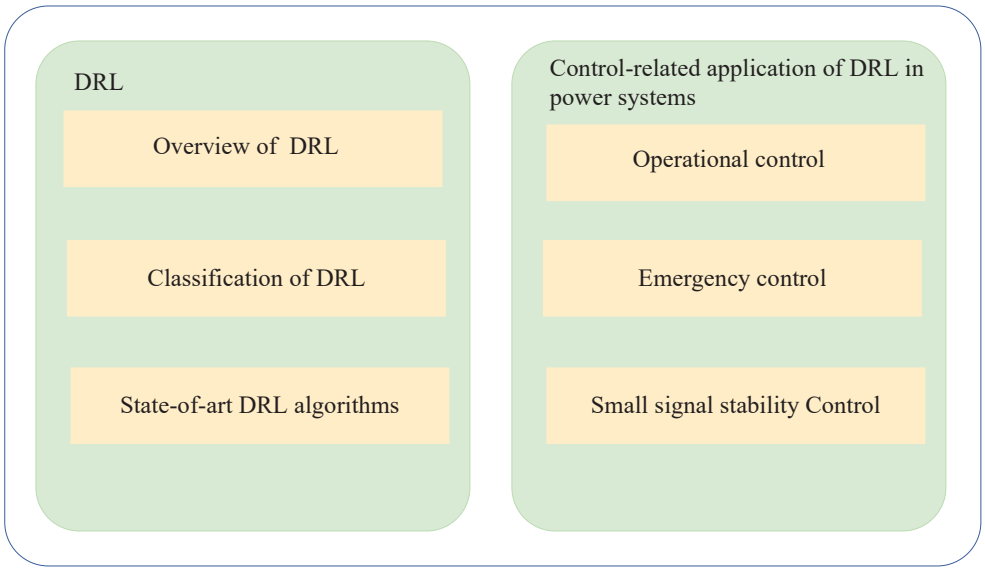

Figure 1. The structure and main content.

2. Review of the Deep Reinforcement Learning Algorithms

The majority of artificial intelligence (AI) issues that occur today are resolved via machine learning, a subfield of AI. Typically, there are three types of machine learning: supervised learning, unsupervised learning, and RL [13]. Supervised learning is generally used to train and improve the learning system for predictions, classifications, or regressions by evaluating the output and confirming the labels of data [14]. There are two underlying assumptions of supervised learning. Firstly, the input data are assumed to be independent. Otherwise, it is difficult for the learning system of supervised learning to improve itself. Secondly, the learning system has been told by the data label to correct its predictions through the actual labels. Unsupervised learning is used to discover the hidden patterns and search for the differences in unlabeled training datasets, and is generally is applied for clustering and reducing dimensions.

2.1. The Fundamentals of Reinforcement Learning

There are two main components of RL. As shown in Figure 2, the agent and the environment. Unlike supervised or unsupervised learning, RL allows the agent to explore the actions with maximum cumulative rewards instead of being told which actions should be taken [15]. Moreover, the agent still needs training in many episodes through trial and error in the environment. In particular, each episode is a trajectory of states, actions, and rewards across time. $(s_0, a_0, r_0, s_1, a_1, r_1, \dots)$ in RL will be terminated when it reaches a certain goal. Moreover, RL is sequential, long-term, and emphasizes on the accumulation of rewards over time as opposed to the immediate rewards of supervised or unsupervised learning.

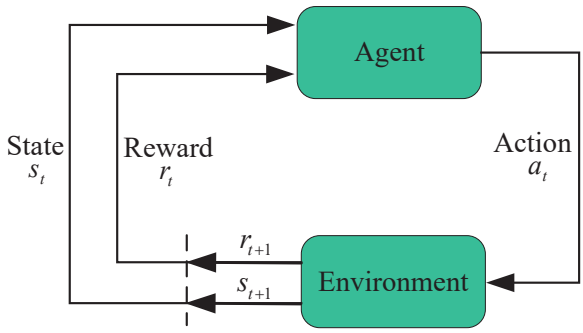

Figure 2. The Interaction between the environment and agent.

The agent is both a decision-maker and a learner. The environment, which is made up of everything around the agent, is the object with which the agent interacts. The agent and environment interact with each other at each of a series of discrete time steps $t = 0, 1, 2, 3, \dots$. At each time step t, the agent obtains some representation of the state S_t of the environment, where $S_t \in \mathcal{S}$ and $\mathcal{S}$ are the sets of potential states. On the basis of that representation, the agent chooses an action A_t, where $A_t \in \mathcal{A}(S_t)$ and $\mathcal{A}$ are the sets of available actions in the current state S_t of the environment [15].

2.2. Markov Decision Process

The issues that need to be solved in RL can be dealt with in an MDP. The purpose of an MDP is to simplify the environment. While the state signal and environment satisfy the Markov property, i.e., the next state and the expected reward must only be decided by the current state and action, it can be formulated as an MDP. An MDP is applied to provide a mathematical framework for describing decision-making problems in a situation where results are both partially unpredictable and partially under the decision maker's control [15]. An MDP can be described as a tuple $\langle S, A, R, P, \gamma \rangle$, where R is the reward

function, P is the transition probability (1) with the transition matrix (2), γ is the discount factor, and $\gamma \in [0, 1)$.

$$P(s_{t+1} \mid s_t, a_t) = P(s_{t+1} \mid s_0, a_0, \cdots, s_t, a_t) \tag{1}$$

$$P = \begin{bmatrix} P(s_1 \mid s_1) & P(s_2 \mid s_1) & \dots & P(s_N \mid s_1) \\ P(s_1 \mid s_2) & P(s_2 \mid s_2) & \dots & P(s_N \mid s_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(s_1 \mid s_N) & P(s_2 \mid s_N) & \dots & P(s_N \mid s_N) \end{bmatrix} \tag{2}$$

The illustration of an MDP is shown in Figure 3. At each epoch of the MDP, the agent takes action a_t based on the current state s_t of the environment, receives reward r_t from the action, and moves to the next state s_{t+1}.

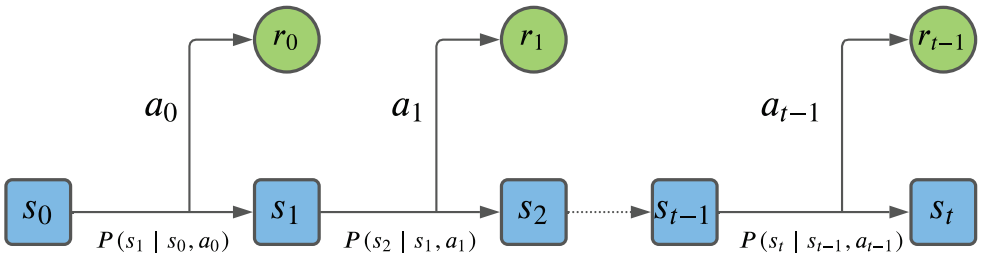

Figure 3. Illustration of the Markov decision process.

The problem is supposed to match the framework of the MDP to support the theoretical result of RL. In addition, RL can still be an acceptable approach even though it slightly deviates from the definition of an MDP. Conversely, suppose the Markov property is not satisfied with the problems, e.g., the partially observable Markov decision process (POMDP). In that case, RL may suffer from non-stationary issues, which results in an inaccurate outcome.

2.3. Value Functions

An essential criterion for judging the action in most RL algorithms is to estimate the value functions of the states or state-action pairs. The state-value function or action-value function evaluates how acceptable the action is for the agent to accept a given state. For an MDP, the state-value function $V_\pi(s)$ can be defined as:

$$V_\pi(s) = \mathbb{E}_\pi[R_t \mid S_t = s] \tag{3}$$

where $\mathbb{E}$ denotes the expected value for the agent by following policy π at any time step t. In addition, R_t is the total discounted return that estimates the value of cumulative future awards:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{4}$$

Furthermore, the state-action function V_π and the action-value function $Q_\pi(s, a)$ can be decomposed into a Bellman equation as:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma V_\pi(s_{t+1}) \mid s_t = s] \end{aligned} \tag{5}$$

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma Q_\pi(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a] \end{aligned} \tag{6}$$

For instance, classic Q-learning uses the action-value function to determine the Q value, the updating rule is shown in an iterative form:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\Big] \tag{7}$$

The optimal policy π^* denotes the policy that obtains the highest cumulative reward in the process. The following are the optimal state-value function and action-value function:

$$\begin{aligned} V^*(s) &= \max_\pi V_\pi(s) \\ Q^*(s,a) &= \max_\pi Q_\pi(s,a) \end{aligned} \tag{8}$$

In RL, the DP applies Equation (5) and determines the current value based on the following state's value (bootstrapping). However, the DP method is required to obtain the model's state transition probability p, and the model must be known. The value function of the DP is:

$$V(s_t) \leftarrow \sum_a \pi(a \mid s_t) \sum_{s_{t+1},r} p(s_{t+1}, r \mid s_t, a)[r + \gamma V_\pi(s_{t+1})] \tag{9}$$

Temporal difference (TD) learning combines the advantages of the DP and model-free Monte Carlo (MC), which uses sampling means instead of expectation. Like the MC, TD learning also learns from the experience but updates the parameter at each time step t instead of at the end of the MC episode. TD learning applies the original value function and single return to determine the target value, which can be expressed as (10). TD learning is used for predictions and is commonly used for learning approaches, such as Q-learning and Sarsa, for evaluating value functions via bootstrapping in an online, model-free, and entirely incremental manner [16].

$$V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)] \tag{10}$$

where α is the learning rate.

Bootstrapping methods are similar but not gradient descent methods because the objective function depends on the predicted weights. TD learning, as a bootstrapping approach, is easier to learn since it can continuously learn online. In addition, because the objective depends on the predicted weights, bootstrapping methods are not examples of gradient descent. Furthermore, bootstrapping approaches will cause a non-uniform overestimation issue in DRL, which will be discussed in the policy-based algorithm section.

2.4. Classification of Deep Reinforcement Learning Algorithms

DRL algorithms break the barrier and exceed the limits of both RL and deep learning algorithms. Classic RL usually suffers from the curse of dimensionality [17] in various sequential decision-making problems. The number of states generally rises exponentially in proportion to state variables. Moreover, some states and variables in the real environment are continuous and highly dimensional. Therefore, these problems bring many challenges for RL methods to be applied in practice. To address this problem, a DNN is used as the function approximator with extra hidden layers between the input and output layers. As a result, a new representation of the input from the previous layer is typically obtained by a non-linear transformation or activation function, such as a logistic and rectified linear unit (ReLU). Deep learning is built around the concept of distributed representation, which implies that a feature may be the result of multiple inputs, and an input may be used to denote various features. Distributed representations have exponential advantages

over centralized representations, which help RL overcome the problem of the curse of dimensionality. While combing deep learning and RL, DNNs also can be used for the approximation of RL components, such as value functions, $\hat{v}(s;\theta)$ or $\hat{q}(s,a;\theta)$, policy $\pi(a \mid s;\theta)$, where θ is the parameter of the neural network [13]. Therefore, DRL combines deep learning and RL to solve control problems with more high dimensional states and larger state spaces. In general, DRL-based approaches can be classified as value-based and policy-based, as shown in Figure 4.

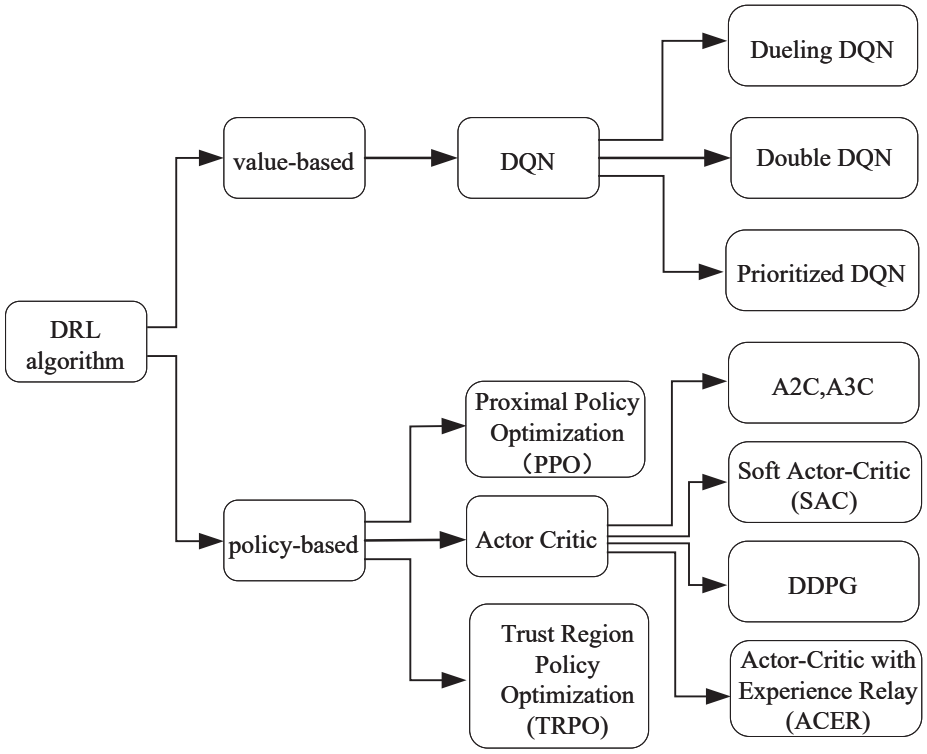

Figure 4. The classification of common DRL algorithms.

2.4.1. Value-Based Algorithm

Almost all model-free RL-based approaches are either value-based or policy-based. The deep Q-network (DQN) is one of the representative value-based methods that ignited the field of RL in 2015 [4]. In the DQN algorithm, a DNN is used as an approximator to fit the action-value function instead of the tabular value in classic Q-learning. The main contributions of the DQN are to build a replay buffer and develop an experienced replay mechanism to store transitions that mitigate the correlations in the training data sampled by a mini-batch. Another contribution is the target network. The target network $Q(s,a;\theta')$ is a clone of the online network $Q(s,a;\theta)$ that has the same network structure but different parameters. Parameter θ of the online network is updated by stochastic gradient descent (SGD), but parameter θ' is kept frozen in each episode and is only updated in a specific period by replacing the value of θ or the weighted average value of θ and θ'. To minimize the difference of the output between the target network and the online network, their parameters can be optimized by minimizing the loss function, as shown in (11).

$$L(\theta) = \mathbb{E}_{\pi}\left[\left(Q(s_t, a_t \mid \theta) - r(s_t, a_t) - \gamma \max_{a} Q\left(s_{t+1}, a_{t+1} \mid \theta'\right)\right)^2\right] \tag{11}$$

Furthermore, action a_t is selected between a random action or the output of $\arg\max_a Q(\phi(s_t), a;\theta)$ by following the ϵ-greedy policy, where $\phi(s_t)$ is sequence s_t.

A double DQN (DDQN) [18] and dueling DQN [19] are two improved examples of DQN algorithms. To address the non-uniform overestimation issues caused by TD learning in a DQN, a DDQN follows the action selection through the greedy policy and

naïve update as the DQN does, but it evaluates the value of the action through the target network. A DDQN alleviates the overestimation issue compared to the naïve update and target network update, although the overestimation issue still exists. Therefore, a DDQN has better performance than a DQN for playing Atari games. A dueling DQN adopts a dueling network architecture to estimate the action-value function $Q(s, a)$ by combining it with the state-value function to achieve faster convergence.

2.4.2. Policy-Based Algorithm

The majority of value-based techniques produce deterministic policies that take the same action under the same set of circumstances. Since the agent employs a certain policy to investigate, it may cause the agent to try insufficient actions to find helpful learning signals. Additionally, value-based methods, such as a DQN, are usually not suitable for handling high-dimensional or continuous action space applications since they need to discretize the action domain, resulting in a partial optimal solution and exponential increase of calculations. To address these issues, policy-based methods, such as a policy gradient (PG), obtain DNNs to produce a stochastic policy $\pi_\theta(a \mid s)$ to estimate the probability of taking action a in given state s by updating parameter θ via gradient ascent methods [20]. In addition, a PG can be applied in either a discrete or continuous action space, which depends on how the policy model is built.

Actor–critic is a policy-based RL approach that blends TD learning and a policy gradient, in which the actor refers to the policy function $\pi(a \mid s)$ and the critic refers to the value function $V_\pi(s)$. The critic network estimates the value function of the current policy to evaluate how 'good' the policy is. The actor–critic algorithm's parameters are updated at each time step rather than at the conclusion of each episode thanks to a characteristic of the value function.

As an extension of the PG, an actor–critic algorithm named the deep deterministic policy gradient (DDPG) [21], which combines deep learning and deterministic policy gradient (DPG), adopts the experience replay mechanism and target network from the DQN and can deal with continuous states and action spaces. To improve the efficiency of exploration, a common approach is to add the correlated time-dependent Ornstein–Uhlenbeck (OU) noise [22] or the uncorrelated Gaussian noise to the action selected by the actor online network. In practice, the Gaussian noise might not be chosen since the Gaussian hyperparameters in the noise process are hard to tune manually and more likely result in sub-optimal policy [23].

To ensure the stability of both actor networks and critic networks, the fixed network is used in a DDPG. Following the training of mini-batched data sampled from the replay buffer, the DDPG usually updates the parameter of the online actor network by the policy gradient shown in (13). For the critic online network, the parameter is updated by the TD algorithm of the loss function shown in (12). Compared to the DQN, the parameter of each online network is updated by the soft target update shown in (14) instead of replacing the value of the parameter at regular intervals.

$$L(\theta^Q) = \frac{1}{N}\sum_i \left(y_i - Q\left(s_i, a_i \mid \theta^Q\right)\right)^2 \tag{12}$$

$$\nabla J(\theta^\mu) \approx \frac{1}{N}\sum_i \nabla_a Q\left(s, a \mid \theta^Q\right)\Big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\Big|_{s_i} \tag{13}$$

where N is the number of transitions (s_i, a_i, r_i, s_{i+1}) sampled in one batch; θ^μ and θ^Q are, respectively, the parameters of the actor and critic online networks; $\theta^{\mu'}$ and $\theta^{Q'}$ are, respectively, the parameters of the actor and critic target networks.

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \end{aligned} \tag{14}$$

where τ is a constant $\ll 1$ and is usually set as 0.001.

The policy-based RL methods usually provide more excellent convergence guarantees than value-based RL methods [24], especially when using neural networks to approximate functions, which can handle larger scales of state and action spaces [25].

2.5. Multi-Agent Deep Reinforcement Learning Algorithm

For single-agent DRL methods, the critical issue is that the environment is not stationary from each agent's perspective, while each agent is changing its policy in a multi-agent environment. Thus, the corresponding change in the environment is not explainable for each agent, which causes learning instability issues and discards the experience replay mechanism. However, straightforwardly mapping the single-agent DRL to a multi-agent environment to train each agent separately [26] usually results in no convergence or overfitting issue because each agent carries different policy networks with the distinct policy score function $J(\theta^i)$. Then, the individual agents' long-term rewards are now dependent on the policies of all other agents, which means that while an individual agent updates the policy network, other agents will change their policy network correspondingly.

The multi-agent deep reinforcement learning (MADRL) algorithm is proposed to solve sequential decision-making problems in a multi-agent environment with multiple agents. The behaviors of each agent impact both the state of the entire environment and the reward for each agent. In this case, as an expansion of an MDP, Markov games, also known as stochastic games [27] that deal with the discrepancy among the agents, are required in MADRL algorithms. A Markov game is described as a tuple $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{R^i\}_{i\in\mathcal{N}}, \gamma \rangle$, which can be explained as follows:

- $\mathcal{N} = \{1, \cdots, N\}$ is the set of agents i, where $i \in \mathcal{N}$;
- State space $\mathcal{S}$ is the observation of the state space from all agents;
- $\mathcal{A}^i$ is the action space of agent i;
- $\mathcal{P}$ is the transition probability from s to s_{t+1}, where $s, s_{t+1} \in \mathcal{S}$ for any joint action $a \in \mathcal{A}$, where $\mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$;
- R^i is the reward function that estimates the reward for agent i from the transition of (s, a) to s_{t+1};
- γ is the discount factor and $\gamma \in [0, 1)$.

Accordingly, the value function of MADRL (15) also needs to be modified to adapt to the multi-agent scenarios. At each time step, action a_t^i is taken by agent i based on the observation of environment state s_t. Reward $R^i(s_t, a_t, s_{t+1})$ is gained by agent i as the environment transits to the next state s_{t+1}, while reward R^i is also dependent on another's action a^i. As a result, the value function $V^i_{\pi^i,\pi^{-i}}(s)$ is used to explore the optimal policy $\pi(a_t \mid s_t) := \prod_{i\in\mathcal{N}} \pi_t^i(a_t^i \mid s_t)$ for all joint policies according to state s_t, since the optimal policy of each agent is not only decided by its own.

$$V^i_{\pi^i,\pi^{-i}}(s) := \mathbb{E}\left[\sum_{t\geq 0} \gamma^t R^i(s_t, a_t, s_{t+1}) \mid a_t^i \sim \pi^i(\cdot \mid s_t), s_0 = s\right] \tag{15}$$

where $-i$ denotes all indexes in $\mathcal{N}$ of the agents except agent i.

The most common standard to determine the convergence of MADRL is the Nash equilibrium (NE) [28]. The situation of a stationary NE (16) of a Markov game is that it is impossible for agent i to have a better-then-expected performance by modifying its policy, while the policies of other agents $-i$ remain the same [29]. In other words, once the NE indicates the convergence, agents are not supposed to change their policy because none of the agents has any incentive to deviate.

$$V^i_{\pi^{i,*},\pi^{-i,\epsilon}}(s) \geq V^i_{\pi^i,\pi^{-i,*}}(s) \quad \text{for any } \pi^i \tag{16}$$

Conventional voltage/frequency control approaches can be classified according to the communication modes: centralized, decentralized, local, and distributed control. Centralized control can obtain global optimization but it requires costly and reliable communication

lines, and this is hard to achieve in the current power systems. Similarly, distributed control can coordinate well among control devices but it is also limited to communication conditions. Additionally, local control can respond quickly without extensive communication linkages and only needs the local measurement. However, poor coordination often results in sub-optimal solutions that might not satisfy all constraints required for power systems. An alternative approach is centralized training and decentralized execution (CTDE). By way of illustration of the CTDE framework, an agent can view the other agents' observations and actions as a part of its training, providing a complete picture of the status in the whole system. Agents are expected to consider others and adjust themselves during a decentralized execution. Ref. [30] present an adaptation of actor-critic methods that considers action policies of other agents and is able to successfully learn policies that require complex multi-agent coordination, besides, a training regimen utilizing an ensemble of policies for each agent that leads to more robust multi-agent policies is introduced.

3. DRL for Modern Renewable Power System Control Applications: A Review of Recent Works

The power system control applications are reviewed in terms of operating states: normal operating state and emergency state. Restorative and preventive control is not included in this paper because there are rare considerations in these fields. Furthermore, the considerations under each operating state are reviewed according to the control level and the type of DRL algorithms. The summary table is shown in Table 1.

3.1. Application of DRL in the Operational Control of a Modern Renewable Power System

Voltage deviation is the main criterion to identify the normal operation of the power system. Large voltage deviations can affect the operating efficiency, shorten the life of electrical equipment, and bring safety concerns. It is imperative that the voltage remains within a range around the nominal value. Generally, the voltage magnitude at substations should remain within the normal range around $\pm 10\%$ nominal voltages for the distribution system and $\pm 5\%$ for the transmission system. At present, the common devices used for voltage management in distribution networks are on-load tap changers (OLTCs), shunt capacitors, PV inverters, and static var compensators (SVCs). The author of [31] proposes a DDPG-based approach to adjust tap ratios of OLTCs and determines the optimal policy to maintain the voltage within a safe range while considering minimizing the economic costs in distribution networks. However, the life span of the traditional VVC devices would be drastically reduced under an excessive number of procedures. Moreover, the slow-action devices cannot handle the rapid voltage fluctuation caused by the renewable generations and demand response. In [32], a surrogate-model-enabled DDPG-based approach is presented to control PV inverters, SVCs, and the active power curtailment of PV inverters to solve the voltage fluctuation caused by PV generations in real-time. The surrogate model learns the optimal control policy by interacting with a DRL agent in a supervised manner to represent the non-linear mapping relationship of the power injection and the voltage fluctuation of each node.

The voltage regulation devices can be divided into slow and fast regulation devices. Conventional devices, including static capacitor banks (CBs) and OLTCs, are mechanical devices that are sluggish to respond to changes in voltage (e.g., seconds or minutes) [33]. While the static var compensators (SVCs) and static synchronous compensators (STATCOMs) have a faster reaction time (e.g., seconds or under a second) and better reactive power injection capability as an alternative. In this regard, the voltage regulation can be broken down into multi-timescale optimizations because of the different reaction times and characteristics of mechanical and power electronic devices. The author of [34] decomposes voltage control in distribution grids in two timescales and applies both physical-driven and data-driven-based methods to the voltage control problem. For the former, the precise or approximate grid models are used to obtain the optimal setpoints for inverters in a fast timescale. The latter applies the DQN algorithm to determine the reactive power control

strategy of switching the shunt capacitors in a slow timescale (e.g., hours, days). Moreover, the author of [35] applies hybrid DRL algorithms to different timescale control devices to generate optimal control policy in both fast and slow timescales with the continuous and discrete domain. In particular, multiple agents are divided into DQN-based and DDPG-based agents, which are used for discrete actions, including the configuration of capacitors in a slow timescale and continuous actions, such as the control strategies of PV inverters and energy storage batteries, in a fast timescale. Unlike the DQN, the DDPG can directly handle continuous state-action spaces instead of discretization. Furthermore, the two types of agents work collaboratively to produce real-time voltage control strategies. The author of [36] designs a DRL-based two-stage volt/var (VVC) architecture that coordinates voltage-regulating devices in real-time while reducing the power loss. In the first stage, the scheduling of the OLTCs and CBs are sent a day ahead from the hourly PV and load predictions using optimal power flow (OPF) in the central controller. Then, in the second stage, the dispatch results of the OLTCs and CBs are regarded as the input of the MADDPG algorithm to learn and explore the optimal reactive power point of PVs using PV inverters in a fast timescale. To alleviate voltage violations caused by the uncertainty of EVs and load in the active distribution network (ADN), the author of [37] applies the DDPG to a two-stage method to alleviate voltage violation. To extend the previous work, reactive power dispatch is also sent a day-ahead to lessen the power losses using the mixed-integer second-order cone programming (MISOCP) in the first stage. Then, the DDPG-based method is applied at each charging station (CS) for voltage control within an acceptable range in the second stage.

A single objective for optimization might be insufficient to satisfy the required goals of the sequential decision-making process in dynamic systems. While dealing with these situations, the MDP of DRL-based VVC approaches is replaced by the constrained Markov decision process (CMDP) [38] to achieve the objective with constraints, which also helps comprehend the trade-off between several purposes. The author of [39] formulates the VVC problem as a CMDP and introduces a high-efficiency off-policy constrained soft actor–critic (CSAC)-based approach to provide optimal VVC strategies with operating cost constraints in the distribution network to satisfy the high-security requirements. Likewise, the author of [40] proposes a safe DDPG-based DRL approach to achieve optimal voltage control by coordinating multiple hybrid distribution transformers (HDTs) with minimal power losses in the formulation of a CMDP, where data are collected from the HDT sensors and ADN. The physical constraints are illustrated with the applicable range of the available reactive power of an HDT. The formulation of the CMDP combines the objectives of keeping all bus voltages within acceptable limits while reducing power losses in the environment of the ADN. In particular, the definition of 'safe' is reflected in the safety layer at the top of the DDPG's actor network, which helps correct the control strategies and ensure that the bus voltage does not exceed the acceptable range. However, it is worth noting that HDTs are considered as the significant control devices in [40], which are not equipped in most of the existing distribution networks.

Apart from the control applications at the distribution network control level, the system-wide control is more complex for decision-making. The author of [41] designs the DQN-based autonomous voltage control (AVC) framework 'grid mind' for the power grid voltage control in a steady state. Specifically, the 'grid mind' framework is separated into offline and online training. In particular, agents can adapt to the environment and explore optimal control actions based on experience in offline simulations. In the online session, the agent implements the control action based on the real-time data collected from SCADA or directly from PMUs, and the supervisor will verify the submitted control policy. Additionally, the author of [42] extends the work of the 'grid mind' framework for the AVC in the power grid and applies both the DQN and DDPG to train DRL agents in different schemes. The DDPG is used for the problems with continuous action space, such as controlling the voltage set-points of generators, and the DQN is applied for the issues with discrete action space, such as switching shunt capacitors.

With the increasing scale of power systems, it is inappropriate for single-agent DRL algorithms to regard the whole power system as a single region for calculations. It should be proportionally distributed to target regulation units. However, agents increased exponentially along with the dimension of state and action spaces, known as the 'curse of dimensionality'. The recent research shows that the MADRL-based methods with composite function approximation can improve scalability. The MADRL-based approaches for voltage control at the distribution network level are introduced below. The author of [43] applies the multi-agent soft actor–critic (MASAC) algorithm to the approach in centralized training with a surrogate model and decentralized execution manner, which coordinate both the battery storage system (BSS) and SVCs to decrease the voltage deviation while keeping a low degree of active power curtailment for PVs. In particular, a sparse pseudo-Gaussian process (SPGP) replaces the original power flow model to determine the reward function in the training process. The SPGP is equivalent to the power flow equations in terms of the input–output relationship to learn about the mapping between node voltage magnitude and both reactive and active power injection but it requires fewer measurements. Similarly, it is necessary to apply a network partition before applying the MADRL algorithm to voltage control problems. The entire network is divided into several sub-regions according to active and reactive voltage sensitivity as well as to electrical distance. The centralized training manner guides all agents to achieve coordinated control using the SPGP mode. As opposed to a single cumulative bonus in common DRL algorithms, the SAC actor functions maximize the sum of the expected rewards and entropy to encourage agents to explore more in the training process. The author of [44] explores an optimal VVC control strategy that coordinates the reactive power of SVCs and PV inverters in sub-networks in an ADN based on the multi-agent twin delayed deep deterministic policy gradient (MATD3) algorithm integrated with an attention model based on the CTDE framework. The optimal partition results of the ADN are determined using spectral clustering, which is an unsupervised learning technique derived from spectral graph theory based on voltage-reactive power sensitivity. The improved MATD3 algorithm with an attention model is used to allow each agent to pay more attention to the detailed information that is primarily concerned with the reward. Additionally, compared to the other decentralized and distributed methods, the proposed method only needs local information, and no communication among agents is required in a distributed manner. The author of [45] designs a two-timescale multi-agent voltage control framework to coordinate different layers of agents for the ADN. Agents are categorized into two levels, upper-level agent and lower-level agent. For instance, the upper-level agent takes the ADN's global states into account to decrease the voltage variation and minimize long-term switching numbers of OLTCs and SCs by the SAC algorithm in the slow time-scale control. In addition, the upper-level agent's actions are sent to the lower-level agents for the coordinated management of PV inverters at each time step by the MASAC algorithm in fast time-scale control instead of pre-determination by the stochastic programming method. Moreover, the lower agent is responsible for network partition based on the voltage-reactive power sensitivity. The two-level agents are then trained simultaneously while exchanging information from the reward signal computed by the surrogate model to establish the systematic coordination among different assets. In [46], a robust regionally coordinated VVC (RRV-VVC) approach is described in a multi-agent context. The RRC-VVC approach maintains the coupled power flow linkages while dividing the distribution network into numerous sub-networks. In order to regulate voltages while reducing power losses in the presence of spatial and temporal uncertainty in PV power generation and loads via PV inverters, the MADDPG is applied to each sub-network based on the CTDE architecture. Because loads and renewable energy generation vary according to location and are subject to short-term intermittency and volatility, stochastic programming is used to account for the spatial and temporal uncertainty. Similar to the previous literature, voltage deviation and power loss are considered in the reward function by the classic weight sum algorithm. Therefore, the multi-objective functions are converted into a single-objective function.

There are also considerations of the micro-grid level using MADRL-based methods for voltage control. In [47], the PowerNet algorithm, which is a decentralized on-policy MADRL approach, is presented to achieve secondary voltage control (SVC) in the micro-grid. The proposed communication protocol allows each agent to communicate with its neighbors for the required data, such as the encoded information of states, to boost learning effectiveness. Then, a spatial discount function is introduced to describe the physical distance's correlations between different agents, which is included in the reward function to solve the problem of instantaneous global reward, such as the credit assignment problem. Moreover, to address the noise problem caused by the on-policy RL approach, the sampled stochastic actions are smoothed by PowerNet to mitigate the action fluctuations to ensure the desired performance.

The OPF-based approach has been widely applied to support system-wide voltage regulation in power systems by utilizing the convex relax methodology [48] to tackle non-linear and non-convex issues [49]. Nevertheless, single-point failure, communication delay, and scalability concerns may distort the results and affect the performance of the real-time control of OPF-based methods. The partition-based distributed coordination control using MADRL methods is a trendy alternative approach in recent research [50–52]. Since the standards of the communication condition required by the MADRL approaches in most scenarios are relatively low, which are appealing to the current limited communication conditions in today's power system, they mitigate the high-cost deployment of communication devices. The MADDPG is refined by assigning an individual replay buffer to each agent and is applied to the multi-agent AVC (MA-AVC) scheme in [53]. The power grid is first segmented into several partitions roughly according to geographic considerations, and an agent is assigned to each partition with an individual actor, critic, coordinator, and replay buffer. In addition, the partition remains until it is guaranteed that there are no uncontrollable buses. This process is called the post-portion adjustment. Furthermore, the proposed MA-AVC scheme ameliorates the problems caused by delayed communication since only some specific data are needed to be shared among agents, such as the next estimated action value after execution in the training process. The author of [54] adopts an optimal steady-state voltage regulation in the urban power grid. The MADDPG associated with the dynamic reward function is applied to solve the power flow equations to minimize the voltage deviation, and N-1 contingencies are also taken into consideration. There are two types of agents in the control scheme, SVC and system agents. The SVC agent calculates active power injections based on local data and the setpoint by the DDPG algorithm, and the system agent provides the available voltage setpoint for the SVC agent using the MADDPG algorithm. They are coordinated to achieve voltage regulation. Specifically, the DDPG is applied to train the SVC agent to determine the active power injections based on the local measurement and the voltage setpoint is decided by the system agent with the MADDPG.

Active and reactive power must be balanced between generation and consumption, which is frequently accomplished by employing centralized control centers equipped with automated generation control (AGC) capabilities to manage the different generating units in secondary frequency/voltage control. Load frequency control (LFC) is included in AGC and refers to the active power and frequency regulation. LFC is used to fulfill a region's local demands first and later to reduce the steady-state frequency Δf to zero with a fast response in only a few seconds to maintain the system stability. The author of [55] introduces an MADDPG approach to regard each generation unit as an agent to perform primary and secondary frequency control in a decentralized manner. The author of [56] presents an MADDPG-based method for the optimal coordinated control by load frequency controllers to adjust the power generation through multiple areas. In addition, an LFC database is established during the initialization process to reduce the training time of the parameter of the DNN based on stacked-denoising auto-encoders (SDAEs), where the training data are collected with labels from area control error (ACE) signals and fine-tuned PID controllers.

The Federal Energy Regulatory Commission (FERC) proposed the performance-based frequency regulation market [57] to maximize the control performance while reducing regulatory mileage in 2011. A secondary frequency regulation encourages fast-responding regulation units, such as wind turbines, photovoltaic power plants, and flexible loads. The author of [58] designs an intelligent AGC (IAGC) framework for proportional–integral (PI) controllers to achieve multi-area comprehensive optimal frequency regulation by adjusting its coefficient in each area and satisfying the performance-based frequency regulation market mechanism and coordinating the dispatch and control algorithm. Specifically, the parameter of each PI controller is trained from the local resources, and tuners are added in each controller for applying the guided-exploration multi-agent twin-delayed DDPG (GE-MATD3) algorithm to achieve the multi-area coordination of the frequency regulation in real-time. The author of [59] applies the curriculum multi-agent DDPG (EIC-MADDPG) algorithm to the distributed intelligent coordinated AGC (DIC-AGC) framework in a CTDE manner for the optimal controller coordination of a multi-area integrated energy system (IES) in the performance-based frequency regulation market. The adoption of the CTDE mechanism only requires the agent in each area of the IES to observe the local state instead of the global state in the whole system. The author of [60] develops a multi-agent distributed multiple-improved DDPG (MADMI-TD3) algorithm based on the virtual generation alliance AGC (VGA-AGC) framework to achieve the coordination of the control algorithm and dispatch algorithm in the performance-based frequency regulation market. In particular, the VGA-AGC framework is mapped to a pyramid structure. To assort the agents into five levels, starting from the top: king agent, general agent, lord agent, and knight agent. More specifically, the task of the agents at the lower level is to execute the generation command from the upper-level agent. The king agent plays the role of the traditional PI controller to observe the global state of the power grid and adjusts the action accordingly.

3.2. Application of DRL in the Emergency Control of a Modern Renewable Power System

At present, two major problems (frequency and voltage instability) have been considered in the existing work so far. Although in past decades, many RL-based techniques have considered the instability problems, including transient angle instability, these are currently rarely considered for DRL-based methods.

The fault-induced delayed voltage recovery (FIDVR) refers to the unexpected delay of several seconds in the voltage recovery after the fault is cleared. It is confirmed that the root cause of the FIDVR is the stalled air-conditioner (A/C) units powered by single-phase induction motors. The goal of emergency control for FIDVR issues is to restore the voltage in order to meet the voltage recovery criterion while shedding the load as little as possible. In [61], a DQN algorithm-based DRL technique to handle the FIDVR issue of under-voltage load shedding (UVLS) is considered the emergency control strategy. The UVLS swiftly relays the lower load demand to the substations when monitored bus voltages fall below predetermined voltage thresholds. In addition, the author of [61] develops reinforcement learning for a grid control (RLGC) open-source platform, and its RL module is built on OpenAI Gym for developing, training, and testing RL algorithms in power system simulators. In [62], a derivative-free DRL algorithm called the parallel to augment random search (PARS) and tailored for the power system is developed to reduce voltage instability by UVLS. In particular, simultaneous perturbation stochastic approximation is used to explore the parameter space of the ARS for more efficient explorations. Moreover, a deep meta-reinforcement learning (DMRL) technique is presented in [63] for emergency voltage controls to quickly adapt to the environment of new power grids. The PARS and meta-strategy optimization (MSO) [64] algorithms are specifically combined in DMRL, allowing the agents to quickly adapt to the new environment of power grids, including power flow circumstances and dynamic parameters, by learning the latent context from prior learning. Considering that the decision-making process of DRL algorithms is usually regarded as black-boxes, a policy extraction framework is proposed in [65] to convert a challenging

DRL model into an understandable UVLS policy. The approach proposed in [66] adopts an event-based MDP for intelligent load shedding, and it also incorporates the knowledge of removing negative and repetitive behaviors to increase the effectiveness of training and decision-making. The author of [67] provides an off-policy soft actor–critic architecture with automated entropy adjustment termed SAC auto-discrete for UVLS to enable efficient and adaptive discrete actions for online emergency voltage control against FIDVR.

Once a system has survived fast transient processes, an imbalance between generation and load demand causes frequency instability. The objective of power system emergency frequency control (PSEFC) is to quickly restore the frequency to an acceptable level of the power system after large power disturbances. In [68], a (D)RL-based PSEFC framework is designed for the operator to make flexible selections depending on the demand of various situations. It is worth noting that the PSEFC framework aims to employ load shedding techniques to maintain the system frequency within an acceptable range after a major power disturbance. Moreover, the PSEFC framework provides four (D)RL algorithms, including multi-Q learning, single-agent Q-learning, multi-agent Q-learning, and DDPG algorithms to help operators decide on emergency scenarios. In particular, the RL-based approaches learn the frequency control policy in offline emergency scenarios with minimal costs. The RL-based approaches apply the corresponding control policy learned from the offline scenario in the online scenarios, which best matches the current emergency scenario. It is important to note that the RL-based techniques have quick speeds but low precision. In contrast, the DDPG algorithm-based approach can handle continuous emergency frequency control in multiple scenarios. There is currently no interrelationship between AGC control strategy and emergency control strategy, as they are analyzed and modeled independently [69]. In this case, AGC cannot respond swiftly, resulting in serious risks to the system and the unit when the system depends solely on the AGC control strategy for frequency stabilization. The author of [70] develops a warm agent exploration distributed multiple delayed deep policy gradient (SAE-MD3) algorithm used for AGC dispatch in the wide-area AGC (WA-AGC) framework to achieve emergency control based on real-time measurement data collected from WAMSs while satisfying the performance-based frequency regulation market. In particular, the SAE-MD3 algorithm is improved from the DDPG algorithm to mitigate the overestimation problem. The WA-AGC framework is divided into four intervals with different starting conditions according to the frequency status of: emergency AGC (EAGC) ($\Delta f > 0.5$ Hz or the emergency control device is activated), conventional AGC (CAGC) ($\Delta f < 0.5$ Hz), AGC transition ($\Delta f < 0.125$ Hz and $\Delta f_{1\text{min}} < 0.05$ Hz within 25 control intervals), and OPF ($\Delta f < 0.125$ Hz and $\Delta f_{1\text{min}} < 0.05$ Hz after 25 control intervals).

3.3. Application of DRL in the Small Signal Stability Control of a Modern Renewable Power System

A DDPG algorithm-based agent is created in [71] for the virtual synchronous generator (VSG) to synergistically alter the rotor inertia and damping coefficient in order to enhance the system's transient performance and small signal stability. Low-frequency oscillation (LFO) with a frequency oscillation range of 0.1–2.5 Hz has grown to be a significant issue in modern renewable power systems as a result of the development of the interconnected power grid. In [72], to successfully dampen LFO, a novel sparsity-promoting adaptive control method for online self-tuning of the PSS parameter settings is proposed, and a DDPG is used to train an agent to learn the sparse coordinated control strategy of the multi-PSS. According to recent studies, the system experiences ultralow-frequency oscillation (ULFO) with a frequency below 0.1 Hz, as a result of improper hydraulic governor settings. In [73], a brand-new dual-branch (DB) parallel damping controller is proposed. A multi-agent DRL (MADRL)-based framework, namely a MATD3-enabled collaborative adaptive control framework, is constructed for the decentralized self-tuning of multi-controllers in order to guarantee the robustness of the proposed controller. Although there is a considerable body of literature on wide-area damping control (WADC), creating a wide-area controller based on large models is still computationally difficult. As a result,

a collection of scalable adaptive dynamic programming (ADP)-based wide-area control schemes that are driven solely by real-time measurements of the system states or outputs using reinforcement learning (RL) is proposed in [74]. In [75], a faster exploration-based DDPG algorithm is proposed to timely dampen oscillations, such as LFO, even under various kinds of uncertainties.

Table 1. Literature summary.

Reference	Control State	Field	Algorithm	Agent Type	Objective
[31]	Normal	Distribution Network	DDPG	Single-agent	Voltage Profile and Economic Cost
[32]	Normal	Distribution Network	DDPG	Single-agent	Voltage Regulation
[34]	Normal	Distribution Network	DQN	Single-agent	Voltage Regulation
[35]	Normal	Distribution Network	DDPG, DQN	Single-agent	Voltage Regulation
[36]	Normal	Distribution Network	DQN, DDPG, MADDPG	Single/ Multi-agent	VVC
[37]	Normal	Distribution Network	DDPG	Single-agent	Voltage Control and Power Loss
[39]	Normal	Distribution Network	CSAC	Single-agent	VVC
[40]	Normal	Distribution Network	DDPG	Single-agent	Voltage Control and Power Loss
[41]	Normal	Micro-Grid	DQN	Single-agent	AVC
[42]	Normal	Operational Control	DDPG, DQN	Single-agent	AVC
[43]	Normal	Distribution network	MASAC	Multi-agent	Voltage Deviation and Power Loss
[44]	Normal	Distribution Network	MATD3	Multi-agent	VVC
[45]	Normal	Distribution Network	MASAC	Multi-agent	Voltage Deviation and Operating Cost
[46]	Normal	Distribution Network	MADDPG	Multi-agent	VVC and Power Loss
[47]	Normal	Micro-Grid	PowerNet	Multi-agent	SVC
[53]	Normal	Operational Control	MADDPG	Multi-agent	AVC
[54]	Normal	Operational Control	MADDPG	Multi-agent	Voltage Deviation
[55]	Normal	Operational Control	MADDPG	Multi-agent	LFC
[56]	Normal	Operational Control	MADDPG	Multi-agent	Multi-area LFC
[58]	Normal	Operational Control	IGE-MATD3	Multi-agent	Multi-Area AGC and FR Mileage Payment
[59]	Normal	Operational Control	EIC-MADDPG	Multi-agent	Multi-Area IES AGC and FR Mileage Payment
[61]	Emergency	Emergency Control	DQN	Single-agent	FIDVR
[62]	Emergency	Emergency Control	PARS	Single-agent	FIDVR
[63]	Emergency	Emergency Control	DMRL	Single-agent	FIDVR
[66]	Emergency	Emergency Control	knowledge-enhanced DRL model	Single-agent	FIDVR
[67]	Emergency	Emergency Control	SACAuto-Discrete	Single-agent	FIDVR
[68]	Emergency	Emergency Control	DDPG, Multi-Q learning	Single-agent	PSEFC
[70]	Emergency	Emergency Control	SAE-MD3	Multi-agent	WA-AGC and FR Mileage Payment

Additionally, the technological modernization of power grid infrastructure has gradually transformed modern power systems into cyber-physical systems. Therefore, cyber security is considered as one of the control-related aspects. Recent studies have shown that cyberattacks can mislead the system operator to perform the wrong operations based on modified incorrect observations of information, resulting in huge economic losses. The author of [76] proposes a DQN-based cyber security assessment to identify the most critical component of the target system that can be used by adversaries to attack, and to mitigate the costs of finding the optimal attack transition compared to random transition policy. In addition, a transition graph is generated using the improved common vulnerability scoring system (CVSS) to assess the complexity of each possible attack path considering various adversarial methods. The author of [77] builds an MDP to describe the defensive procedure against data integrity, and applies DQN detection (DQND) to prevent data integrity attacks. Once the attacker hacks into the SCADA system, except for the long-term attack

on the power infrastructure, it is also possible to trip all transmission lines connected to the substation, which causes serious Nk contingency. In [78], a DDPG is used for the recovery strategy following a cyber attack in order to generate optimal recovery actions shortly after the attack is detected, thus alleviating cascading outage risks. The DDPG-based method can adapt different cyber-attack scenarios and keep exploring the (near-)optimal policy.

Another potential aspect is to utilize CV algorithms to extract the information of grid topology and transfer the knowledge to DRL agents to adapt to the fast varying topology. The author of [79] presents a GCN-DDQN method where a graph convolutional network (GCN) [80] is the combination of a convolutional neural network (CNN) and a graph neural network (GNN) and is used to implement the load-shedding strategy in order to address the short-term voltage stability issues caused by FIDVR, while adapting to the varying grid topology. In particular, the grid topology information is embedded with the node features using the GCN to better capture the topologies and spatial correlations among node features.

4. Discussion and Future Directions

In the deployment of the communication infrastructure and computational ability, RL, as an alternative solution for power system control applications, can be more efficiently used as a model-free machine learning method. In past decades, much of the literature has applied RL algorithms, such as Q-learning and SARSA, to power system applications. The concept of DRL (DQN) was first proposed in 2013, and a batch of new DRL algorithms emerged after years of research and practice. In recent years, DRL-based approaches have been applied to power system applications. As the combination of RL and deep learning, DRL has a greater capacity for feature extraction and generalization than basic RL-based methods. However, there are still many limitations and issues that need to be considered. Based on the reviewed literature, the discussion and future directions are listed as follows:

- Security. Power systems have a high standard of security requirements to guarantee the normal operations of modern society. In the existing DRL works, a standard method for some DRL-based approaches is to formulate the physical and operational constraints as the penalty term and add them to the reward function. Some attempts of safe off-policy DRL-based approaches [39,40] formulate the power system as a CMDP to obtain a constrained policy optimization by taking the physical and operational constraints into account. When the state reaches the boundary of the safety region, particular action will be taken to drive it back. The state can still be outside the safety boundary due to these methods' 'soft' manner. Therefore, it is still hard to identify whether the well-trained control policy is safe and completely abides by all possible constraints in the real-world system;
- Scalability. The large-scale power systems are more complex and provide more operational actions and conditions for DRL-based approaches to consider. For single-agent DRL algorithms, as the number of agents increased, the dimension of action and state spaces also grew exponentially. This phenomenon is known as the 'curse of dimensionality', especially for DQN-based approaches. Additionally, the single-agent DRL algorithm uses the centralized framework, which cannot handle the communication burden in large-scale power systems. Compared to the single-agent DRL-based approaches, the MADRL algorithms are helpful to improve scalability. In the existing work, it was observed that the largest scale of test systems was the IEEE 300-bus system. However, since MADRL-based approaches have not been applied to practical large-scale power systems yet, its performance is still needed to be verified in the future;
- Data quantity and quality. Sample complexity is the number of training samples required by a machine learning method to learn a target function successfully. In this case, the larger the scale of the power system, the greater the sample complexity becomes for the DRL-based approaches to obtain the optimal or near-optimal policy. However, data quantity is one of the critical factors that affect the training speed, but

is not the bottleneck of DRL algorithms because the power grid simulator can generate data efficiently. Compared to the power grid simulator, there are a lot of 'bad data' (BD) in the online measurements of the power grid, including missing, outlier, and noise data, mainly caused by electric and magnetic field (EMF) interference and meter device failures. Hence, preprocessing the raw data from the power grid for the input of DRL-based approaches is necessary. With the deployment of measurement devices, it is permissible to use big data techniques to benefit DRL algorithms. The main existing fields of big datasets in power systems include (1) field measurement, (2) weather data, (3) geographic information system (GIS) data, and (4) market data [81]. Big data platforms and data mining approaches can increase situation awareness (SA), data processing, event clustering, classification, and detection. In particular, the SA system monitors the power systems for the heterogeneous data from the SCADA system or the installed intelligent electronic devices (IEDs) and PMUs, and identifies potential states, such as voltage drop, transient oscillation, and line tripping [82]. The data processing, such as wrangling and dimensionality reduction of big data in power systems, can improve the data and computation efficiency [83]. In addition, the clustering, classification, and detection of events can enhance the scalability and data efficiency of DRL algorithms;

- Efficiency. It is observed that many widely used actor–critic algorithms, such as a DDPG and PPO, suffer from long-time training and hyperparameter tuning in power system applications. In particular, each gradient step requires the generation of new samples according to the latest policy, requiring extensive training and tuning time. Even small-scale power systems take more than days or weeks to obtain a well-trained control policy. Otherwise, performance will be affected. In contrast, the popular value-based DRL methods, such as DQN, are more efficient for the lower-dimensional state and action spaces. However, the increasing size of the power system will cause the curse of dimensionality, which makes computing the optimal policy impossible. Many other attempts have been made to improve the efficiency of the algorithms applied to power systems. The author of [31] employs the Q-value of each possible action to replace the action as the input in order to improve the learning efficiency. The author of [47] introduces the spatial discount function into the reward function to solve the slow learning efficiency problem of MADRL caused by the credit assignment problem of instantaneous global reward design. The author of [56] generates the LFC database of the response data from the PID controller to train the DNN's parameters through supervised learning, which requires less time because it is not necessary to employ a DDPG for explorations. The author of [62] proposes the PARS algorithm to significantly reduce the number of manual tuning hyperparameters, and the parallelism of power grid dynamic simulations accelerates the training speed;
- Parameter tuning. Many existing algorithms generally have more than 20 hyperparameters, such as the learning rate, the weighted factors, and the penalty factors. Operators have to tune these hyperparameters manually to ensure the desired performance. This is mainly based on experience, which is quite unfriendly to beginners. This is also a known challenge and an active research topic within RL communities. There are some attempts using AutoRL [84,85], which combines DRL and the gradient-free automated hyperparameter optimization, to replace manual and complicated hyperparameter tuning. Additionally, the alternative direction could be to reduce the number of manual tuning hyperparameters. For instance, the PARS algorithm proposed in [62] reduces the number of manual tuning hyperparameters to five;
- Practical ability. In contrast to computer games, it is impossible to repeatedly generate many operating experiences in the actual power systems. For safety concerns, most DRL-based approaches assume using high-fidelity simulators or accurate environmental models for simulating system dynamics and responses, and run offline training in order to avoid the hazards of unsafe explorations in the real world. Therefore, the simulator's accuracy can affect the actual performance of the DRL-trained control

policy in the power system due to the gap between the simulator and the actual system. Additionally, the DRL-based approaches need to use the randomization technique [86] to apply the trained control policy to the different environments of power systems. This might not work well in the rapidly changing power grid environment due to the unexplained generalization capability of the DNNs. In fact, DRL has been applied to some small-scale projects conducted by the China Southern Power Grid (CSG) in recent decades but with undesirable performance [70]. Therefore, it requires further research to consider the practical capacity of DRL methods;

- Generalization ability. Due to the trial-and-error property of DRL algorithms, it is impossible to apply the well-trained control policy from DRL methods to another grid environment. In the existing works, the previous experience is not applicable for a DRL to be applied to a new or even similar power system. There is an attempt of integrating adaptive algorithms with DRL to increase the speed of training. DMRL is developed to enable DRL agents to adapt to the new environment quickly [63] by learning the latent context from the prior learning;
- Preventive and restorative control. Currently, there is no published literature considering preventive and restorative control based on DRL algorithms. Similarly, there are only a few RL-based approaches in this field. The possible reason is that most control schemes in this field cannot be well formulated as an MDP since they cannot be considered as a dynamic optimization problem. This is still a potential field that needs to be explored in the future.

5. Conclusions

Ultimately, this paper reviewed the DRL algorithm and its control application according to the control states and control levels, and control-related applications were also considered. Compared to RL, deep learning in DRL can obtain attributes, categories, or features of objects from the power systems, and then RL can make control decisions according to this information, which makes up for RL's deficiency in feature extraction. This paper shows that the DRL-based approaches are feasible for the power systems, but it also reveals many limitations from the reviewed literature. Although DRL has rapidly improved sequential decision-making problems in theory, method, and practice in the past few years, further research is highly encouraged to pay attention to the limitations discussed in the previous section. Furthermore, it has to go one step further to apply DRL to the practical scheme. Currently, DRL is encouraged to combine model-based approaches to mutually make up for their deficiencies instead of replacing conventional methods. Some DRL control applications in power systems, such as transfer knowledge and big data techniques, are not included in this paper. However, they can be possible extensions for this paper in the future. We expect to see more extensive research and eventually fill in the blanks in control applications for more control levels and states.

Author Contributions: Conceptualization, T.L. and Q.L.; methodology, T.L., Q.L. and Q.Y.; validation, J.L. and X.F.; formal analysis, H.D.; investigation, Q.Y.; resources, T.L. and H.D.; writing—original draft preparation, Q.L. and Q.Y.; writing—review and editing, Q.L., Q.Y. and T.L; supervision, T.L.; project administration, T.L.; funding acquisition, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the science and technology project of the State Grid Corporation of China grant number 5100-202255377A-2-0-ZN.

Data Availability Statement: Not applicable.

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

Abbreviations

ADN	Active Distribution Network
AGC	Automated Generation Control
AI	Artificial Intelligence
AVC	Autonomous Voltage Control
BSS	Battery Storage System
CBs	Capacitor Banks
CMDP	Constrained Markov Decision Process
CNN	Convolutional Neural Network
CSAC	Constrained Soft Actor–Critic
CTDE	Centralized Training and Decentralized Execution
CV	Computer Version
DDPG	Deep Deterministic Policy Gradient
DMRL	Deep Meta-Reinforcement Learning
DNN	Deep Neural Network
DPG	Deterministic Policy Gradient
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
EMSs	Energy Management Systems
EVs	Electric Vehicles
FIDVR	The Fault Induced Delayed Voltage Recovery
FR	Frequency Regulation
GCN	Graph Convolutional Network
GE-MATD3	Guided-Exploration Multi-Agent Twin-Delayed Deep Deterministic Policy Gradient
HDTs	Hybrid Distribution Transformers
IEDs	Intelligent Electronic Device
LFC	Load Frequency Control
LFO	Low-Frequency Oscillation
MA-AVC	multi-Agent Autonomous Voltage Control
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MADMI-TD3	Multi-Agent Distributed Multiple Improved Deep Deterministic Policy Gradient
MADRL	Multi-Agent Deep Reinforcement Learning
MASAC	Multi-Agent Soft Actor–Critic
MATD3	Multi-Agent Twin Delayed Deep Deterministic Policy Gradient
MC	Monte Carlo
MDP	Markov Decision Process
MGs	Micro-Grids
MISOCP	Mixed-Integer Second-Rrder 363 Cone Programming
MLMVN	Multilayered Neural Network with Multivalued Neurons
OLTCs	On-Load Tap Changers
OPF	Optimal Power Flow
OU	Ornstein–Uhlenbeck
PARS	Augment Random Search
PG	Policy Gradient
PMUs	Phasor Measurement Units
POMDP	Partially Observable Markov Decision Process
PSEFC	Power System Emergency Frequency Control
ReLU	Rectified Linear Unit
RES	Renewable Energy Sources
RL	Reinforcement Learning
SAE-MD3	Agent Exploration Distributed Multiple Delayed Deep Policy Gradient
SARSA	State-Action-Reward-State-Action
SDAE	Stacked-Denoising Auto-Encoders

SGD	Stochastic Gradient Descent
STATCOM	Static Synchronous Compensator
SVC	Secondary Voltage Control
TD	Temporal Difference
ULFO	Ultralow-Frequency Oscillation
UVLS	Under-Voltage Load Shedding
WAMSs	Wide-Area Measurement Systems

References

1. Lachs, W. Area-wide system protection scheme against extreme contingencies. *Proc. IEEE* **2005**, *93*, 1004–1027. [CrossRef]
2. Muir, A.; Lopatto, J. *Final Report on the 14 August 2003 Blackout in the United States and Canada: Causes and Recommendations*; U.S.-Canada Power System Outage Task Force: Ottawa, ON, Canada, 2004.
3. Aien, M.; Hajebrahimi, A.; Fotuhi-Firuzabad, M. A comprehensive review on uncertainty modeling techniques in power system studies. *Renewableand Sustain. Energy Rev.* **2016**, *57*, 1077–1089. [CrossRef]
4. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
5. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
6. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354. [CrossRef] [PubMed]
7. Caicedo, J.C.; Lazebnik, S. Active object localization with deep reinforcement learning. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2488–2496.
8. Kong, X.; Xin, B.; Wang, Y.; Hua, G. Collaborative deep reinforcement learning for joint object search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 1695–1704.
9. O'Kelly, M.; Sinha, A.; Namkoong, H.; Duchi, J.; Scalable, R.T. End-to-end autonomous vehicle testing via rare-event simulation. *arXiv* **2018**, arXiv:1811.00145.
10. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement learning and its applications in modern power and energy systems: A review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. [CrossRef]
11. Zhang, Z.; Zhang, D.; Qiu, R.C. Deep reinforcement learning for power system applications: An overview. *CSEE J. Power Energy Syst.* **2019**, *6*, 213–225.
12. Glavic, M. (Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives. *Annu. Rev. Control* **2019**, *48*, 22–35. [CrossRef]
13. Li, Y. Deep reinforcement learning: An overview. *arXiv* **2017**, arXiv:1701.07274.
14. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
15. Sutton, R.S.; Barto, A.G., *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
16. Tsitsiklis, J.N.; Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Autom. Control* **1997**, *42*, 674–690. [CrossRef]
17. Bellman, R. *Dynamic Programming*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 1957.
18. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
19. Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; Freitas, N. Dueling network architectures for deep reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1995–2003.
20. Lapan, M. *Deep Reinforcement Learning Hands-On: Apply Modern RL Methods, with Deep Q-Networks, Value Iteration, Policy Gradients, TRPO, AlphaGo Zero and More*; Packt Publishing Ltd.: Birmingham, UK, 2018.
21. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
22. Uhlenbeck, G.E.; Ornstein, L.S. On the theory of the brownian motion. *Phys. Rev.* **1930**, *36*, 823. [CrossRef]
23. Xu, T.; Liu, Q.; Zhao, L.; Peng, J. Learning to explore via meta-policy gradient. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5463–5472.
24. Agarwal, A.; Kakade, S.M.; Lee, J.D.; Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In Proceedings of the Conference on Learning Theory, Graz, Austria, 9–12 July 2020; pp. 64–66.
25. Liu, B., Cai, Q., Yang, Z., Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv* **2019**, arXiv:1906.10306.
26. Tampuu, A.; Matiisen, T.; Kodelja, D.; Kuzovkin, I.; Korjus, K.; Aru, J.; Aru, J.; Vicente, R. Multiagent cooperation and competition with deep reinforcement learning. *PLoS ONE* **2017**, *12*, e0172395. [CrossRef]
27. Shapley, L.S. Stochastic games. *Proc. Natl. Acad. Sci. USA* **1953**, *39*, 1095–1100. [CrossRef]
28. Nash, J.F., Jr. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef]

29. Filar, J.; Vrieze, K. *Competitive Markov Decision Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
30. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv* **2017**, arXiv:1706.02275.
31. Toubeau, J.-F.; Zad, B.B.; Hupez, M.; De Grève, Z.; Vallée, F. Deep reinforcement learning-based voltage control to deal with model uncertainties in distribution networks. *Energies* **2020**, *13*, 3928. [CrossRef]
32. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Yu, N.; Huang, Q.; Chen, Z. Model-free voltage control of active distribution system with pvs using surrogate model-based deep reinforcement learning. *Appl. Energy* **2022**, *306*, 117982. [CrossRef]
33. Hu, Z.; Wang, X.; Chen, H.; Taylor, G. Volt/var control in distribution systems using a time-interval based approach. *IEEE Proc.-Gener. Transm. Distrib.* **2003**, *150*, 548–554. [CrossRef]
34. Yang, Q.; Wang, G.; Sadeghi, A.; Giannakis, G.B.; Sun, J. Two-timescale voltage control in distribution grids using deep reinforcement learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2313–2323. [CrossRef]
35. Zhang, J.; Li, Y.; Wu, Z.; Rong, C.; Wang, T.; Zhang, Z.; Zhou, S. Deep-reinforcement-learning-based two-timescale voltage control for distribution systems. *Energies* **2021**, *14*, 3540. [CrossRef]
36. Sun, X.; Qiu, J. Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method. *IEEE Trans. Smart Grid* **2021**, *12*, 2903–2912. [CrossRef]
37. Sun, X.; Qiu, J. A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6852–6863. [CrossRef]
38. Altman, E. *Constrained Markov Decision Processes*; CRC Press, Boca Raton, FL, USA, 1995.
39. Wang, W.; Yu, N.; Gao, Y.; Shi, J. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Trans. Smart Grid* **2019**, *11*, 3008–3018. [CrossRef]
40. Kou, P.; Liang, D.; Wang, C.; Wu, Z.; Gao, L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. *Appl. Energy* **2020**, *264*, 114772. [CrossRef]
41. Diao, R.; Wang, Z.; Shi, D.; Chang, Q.; Duan, J.; Zhang, X. Autonomous voltage control for grid operation using deep reinforcement learning. In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Montréal, QC, Canada, 2–6 August 2019; pp. 1–5.
42. Duan, J.; Shi, D.; Diao, R.; Li, H.; Wang, Z.; Zhang, B.; Bian, D.; Yi, Z. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. *IEEE Trans. Power Syst.* **2019**, *35*, 814–817. [CrossRef]
43. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z.; Blaabjerg, F. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of pvs. *IEEE Trans. Smart Grid* **2021**, *12*, 4137–4150. [CrossRef]
44. Cao, D.; Zhao, J.; Hu, W.; Ding, F.; Huang, Q.; Chen, Z. Attention enabled multi-agent drl for decentralized volt-var control of active distribution system using pv inverters and svcs. *IEEE Trans. Sustain.* **2021**, *12*, 1582–1592. [CrossRef]
45. Cao, D.; Zhao, J.; Hu, W.; Yu, N.; Ding, F.; Huang, Q.; Chen, Z. Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems. *IEEE Trans. Smart Grid* **2021**, *13*, 149–165. [CrossRef]
46. Liu, H.; Zhang, C.; Chai, Q.; Meng, K.; Guo, Q.; Dong, Z.Y. Robust regional coordination of inverter-based volt/var control via multi-agent deep reinforcement learning. *IEEE Trans. Smart Grid* **2021**, *12*, 5420–5433. [CrossRef]
47. 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]
48. Low, S.H. Convex relaxation of optimal power flow—Part i: Formulations and equivalence. *IEEE Trans. Control Netw. Syst.* **2014**, *1*, 15–27. [CrossRef]
49. Molzahn, D.K.; Dörfler, F.; Sandberg, H.; Low, S.H.; Chakrabarti, S.; Baldick, R.; Lavaei, J. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Trans. Smart Grid* **2017**, *8*, 2941–2962. [CrossRef]
50. Li, P.; Zhang, C.; Wu, Z.; Xu, Y.; Hu, M.; Dong, Z. Distributed adaptive robust voltage/var control with network partition in active distribution networks. *IEEE Trans. Smart Grid* **2019**, *11*, 2245–2256. [CrossRef]
51. Chai, Y.; Guo, L.; Wang, C.; Zhao, Z.; Du, X.; Pan, J. Network partition and voltage coordination control for distribution networks with high penetration of distributed pv units. *IEEE Trans. Power Syst.* **2018**, *33*, 3396–3407. [CrossRef]
52. Zhao, B.; Xu, Z.; Xu, C.; Wang, C.; Lin, F. Network partition-based zonal voltage control for distribution networks with distributed pv systems. *IEEE Trans. Smart Grid* **2017**, *9*, 4087–4098. [CrossRef]
53. Wang, S.; Duan, J.; Shi, D.; Xu, C.; Li, H.; Diao, R.; Wang, Z. A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning. *IEEE Trans. Power Syst.* **2020**, *35*, 4644–4654. [CrossRef]
54. Zhang, X.; Liu, Y.; Duan, J.; Qiu, G.; Liu, T.; Liu, J. Ddpg-based multi-agent framework for svc tuning in urban power grid with renewable energy resources. *IEEE Trans. Power Syst.* **2021**, *36*, 5465–5475. [CrossRef]
55. Rozada, S.; Apostolopoulou, D.; Alonso, E. Load frequency control: A deep multi-agent reinforcement learning approach. In Proceedings of the 2020 IEEE Power & Energy Society General Meeting (PESGM), Orlando, FL, USA, 16—20 July 2020; pp. 1–5.
56. Yan, Z.; Xu, Y. A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Trans. Power Syst.* **2020**, *35*, 4599–4608. [CrossRef]
57. Zhang, X.; Xu, Z.; Yu, T.; Yang, B.; Wang, H. Optimal mileage based agc dispatch of a genco. *IEEE Trans. Power Syst.* **2020**, *35*, 2516–2526.

[CrossRef]
58. Li, J.; Yu, T.; Zhang, X. Coordinated automatic generation control of interconnected power system with imitation guided exploration multi-agent deep reinforcement learning. *Int. J. Electr. Energy Syst.* **2022**, *136*, 107471. [CrossRef]
59. Li, J.; Yu, T.; Zhang, X. Coordinated load frequency control of multi-area integrated energy system using multi-agent deep reinforcement learning. *Appl. Energy* **2022**, *306*, 117900. [CrossRef]
60. Li, J.; Yu, T. Virtual generation alliance automatic generation control based on deep reinforcement learning. *IEEE Access* **2020**, *8*, 182204–182217. [CrossRef]
61. Huang, Q.; Huang, R.; Hao, W.; Tan, J.; Fan, R.; Huang, Z. Adaptive power system emergency control using deep reinforcement learning. *IEEE Trans. Smart Grid* **2019**, *11*, 1171–1182. [CrossRef]
62. Huang, R.; Chen, Y.; Yin, T.; Li, X.; Li, A.; Tan, J.; Yu, W.; Liu, Y.; Huang, Q. Accelerated derivative-free deep reinforcement learning for large-scale grid emergency voltage control. *IEEE Trans. Power Syst.* **2021**, *37*, 14–25. [CrossRef]
63. Huang, R.; Chen, Y.; Yin, T.; Huang, Q.; Tan, J.; Yu, W.; Li, X.; Li, A.; Du, Y. Learning and fast adaptation for grid emergency control via deep meta reinforcement learning. *IEEE Trans. Power Syst.* **2022**, *37*, 4168–4178. [CrossRef]
64. Yu, W.; Tan, J.; Bai, Y.; Coumans, E.; Ha, S. Learning fast adaptation with meta strategy optimization. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2950–2957. [CrossRef]
65. Dai, Y.; Chen, Q.; Zhang, J.; Wang, X.; Chen, Y.; Gao, T.; Xu, P.; Chen, S.; Liao, S.; Jiang, H.; et al. Enhanced oblique decision tree enabled policy extraction for deep reinforcement learning in power system emergency control. *Electr. Power Syst. Res.* **2022**, *209*, 107932. [CrossRef]
66. Hu, Z., Shi, Z., Zeng, L., Yao, W., Tang, Y., Wen, J. Knowledge-enhanced deep reinforcement learning for intelligent event-based load shedding. *Int. J. Electr. Power Energy Syst.* **2023**, *148*, 108978. [CrossRef]
67. Zhang, Y.; Yue, M.; Wang, J. Off-policy deep reinforcement learning with automatic entropy adjustment for adaptive online grid emergency control. *Electr. Power Syst. Res.* **2023**, *217*, 109136. [CrossRef]
68. Chen, C.; Cui, M.; Li, F.; Yin, S.; Wang, X. Model-free emergency frequency control based on reinforcement learning. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2336–2346. [CrossRef]
69. Zhang, X.; Tan, T.; Zhou, B.; Yu, T.; Yang, B.; Huang, X. Adaptive distributed auction-based algorithm for optimal mileage based agc dispatch with high participation of renewable energy. *Int. J. Electr. Power Energy Syst.* **2021**, *124*, 106371. [CrossRef]
70. Li, J.; Yu, T.; Zhang, X. Emergency fault affected wide-area automatic generation control via large-scale deep reinforcement learning. *Eng. Appl. Artif. Intell.* **2021**, *106*, 104500. [CrossRef]
71. Xiong, K.; Hu, W.; Zhang, G.; Zhang, Z.; Chen, Z. Deep reinforcement learning based parameter self-tuning control strategy for VSG. *Energy Rep.* **2022**, *8*, 219–226. [CrossRef]
72. Zhang, G.; Hu, W.; Cao, D.; Huang, Q.; Chen, Z.; Blaabjerg, F. A novel deep reinforcement learning enabled sparsity promoting adaptive control method to improve the stability of power systems with wind energy penetration. *Renew. Energy* **2021**, *178*, 363–376. [CrossRef]
73. Zhang, G.; Zhao, J.; Hu, W.; Cao, D.; Kamwa, I.; Duan, N.; Chen, Z. A Multiagent Deep Reinforcement Learning-Enabled Dual-Branch Damping Controller for Multimode Oscillation. *IEEE Trans. Control Syst. Technol.* **2022**, *31*, 483–492. [CrossRef]
74. Mukherjee, S.; Chakrabortty, A.; Bai, H.; Darvishi, A.; Fardanesh, B. Scalable designs for reinforcement learning-based wide-area damping control. *IEEE Trans. Smart Grid* **2021**, *12*, 2389–2401. [CrossRef]
75. Hashmy, Y.; Yu, Z.; Shi, D.; Weng, Y. Wide-area measurement system-based low frequency oscillation damping control through reinforcement learning. *IEEE Trans. Smart Grid* **2020**, *11*, 5072–5083. [CrossRef]
76. Liu, X.; Ospina, J.; Konstantinou, C. Deep reinforcement learning for cybersecurity assessment of wind integrated power systems. *IEEE Access* **2020**, *8*, 208378–208394. [CrossRef]
77. An, D.; Yang, Q.; Liu, W.; Zhang, Y. Defending against data integrity attacks in smart grid: A deep reinforcement learning-based approach. *IEEE Access* **2019**, *7*, 110835–110845. [CrossRef]
78. Wei, F.; Wan, Z.; He, H. Cyber-attack recovery strategy for smart grid based on deep reinforcement learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2476–2486. [CrossRef]
79. Hossain, R.R.; Huang, Q.; Huang, R. Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control. *IEEE Trans. Power Syst.* **2021**, *36*, 4848–4851. [CrossRef]
80. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef]
81. Kezunovic, M.; Xie, L.; Grijalva, S. The role of big data in improving power system operation and protection. In Proceedings of the 2013 IREP Symposium Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid, Crete, Greece, 25–30 August 2013; pp. 1–9.
82. Tu, C.; He, X.; Shuai, Z.; Jiang, F. Big data issues in smart grid—A review. *Renew. Sustain. Energy Rev.* **2017**, *79*, 1099–1107. [CrossRef]
83. Xie, L.; Chen, Y.; Kumar, P. Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis. *IEEE Trans. Power Syst.* **2014**, *29*, 2784–2794. [CrossRef]
84. Chiang, H.-T.L.; Faust, A.; Fiser, M.; Francis, A. Learning navigation behaviors end-to-end with autorl. *IEEE Robot. Autom.* **2019**, *4*, 2007–2014. [CrossRef]

85. Faust, A.; Francis, A.; Mehta, D. Evolving rewards to automate reinforcement learning. *arXiv* **2019**, arXiv:1905.07628.
86. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30.