

# Two-timescale Volt/var Control Based on Reinforcement Learning with Hybrid Action Space for Distribution Networks

Yuan Zhou, Yizhou Peng, Leijiao Ge, Luyang Hou, Ying Wang, and Hongxia Niu

**Abstract**—In volt/var control (VVC) for active distribution networks, it is essential to integrate traditional voltage regulation devices with modern smart photovoltaic inverters to prevent voltage violations. However, model-based multi-device VVC methods rely on accurate system models for decision-making, which can be challenging due to the extensive modeling workload. To tackle the complexities of multi-device cooperation in VVC, this paper proposes a two-timescale VVC method based on reinforcement learning with hybrid action space, termed the hybrid action representation twin delayed deep deterministic policy gradient (HAR-TD3) method. This method simultaneously manages traditional discrete voltage regulation devices, which operate on a slower timescale, and smart continuous voltage regulation devices, which function on a faster timescale. To enable effective collaboration between the different action spaces of these devices, we propose a variational auto-encoder based hybrid action reconstruction network. This network captures the interdependencies of hybrid actions by embedding both discrete and continuous actions into the latent representation space and subsequently decoding them for action reconstruction. The proposed method is validated on IEEE 33-bus, 69-bus, and 123-bus distribution networks. Numerical results indicate that the proposed method successfully coordinates discrete and continuous voltage regulation devices, achieving fewer voltage violations compared with state-of-the-art reinforcement learning methods.

**Index Terms**—Voltage/var control, reinforcement learning, hybrid action space, two-timescale, active distribution network.

## I. INTRODUCTION

**V**OLT/VAR control (VVC) is an effective strategy for mitigating voltage violations and has been seamlessly

integrated into distribution management systems, enhancing the voltage distribution efficiency and reducing the network loss [1]. In traditional distribution networks, the implementation of VVC is relatively straightforward, involving the use of voltage regulation tools such as on-load tap changers (OLTCs) and reactive power compensation capacitors to address voltage violations [2]. However, the integration of distributed photovoltaic (PV) systems into active distribution networks presents new challenges: the intermittency of PV generation can lead to high-frequency voltage fluctuations and significant voltage violations [3]. Traditional voltage regulators, which typically operate at a lower frequency, struggle to adapt to these abrupt changes. To maintain the safety and stability of the distribution network in this dynamic environment, a voltage regulation approach must synchronize the operation of reactive power devices such as PV inverters, which operate on a faster timescale, with traditional voltage regulation devices that function on different timescales. This synchronization is essential for effectively managing voltage fluctuations.

In traditional VVC scenarios, OLTCs and switchable capacitor banks (CBs) are commonly employed to achieve effective voltage control. Various model-based VVC methods have been developed to coordinate the dispatch of these devices in traditional VVC setups, including branch-and-cut method [1], oriented discrete coordinate descent method [4], nondominated sorting genetic algorithm [5], and fuzzification method [6]. These methods determine the optimal hourly discrete setpoints for the voltage regulation devices by solving the optimal power flow problem. However, traditional model-based VVC methods heavily rely on accurate and comprehensive distribution network models that include parameters such as topology and line parameters [7]. Obtaining precise and detailed model parameters to support mathematical optimization is often challenging. In contrast, deep reinforcement learning (DRL) based control methods utilize simulated operational data or historical data during the training phase and make decisions based solely on the state of the distribution network during the inference phase. These methods reduce dependence on precise topology and parameter information, significantly improving control efficiency and presenting a promising alternative [8]. As a result, several DRL-based VVC methods have been proposed. For instance, [9] enhances DRL by incorporating a supervised learning model

Manuscript received: June 18, 2024; revised: October 8, 2024; accepted: December 2, 2024. Date of CrossCheck: December 2, 2024. Date of online publication: January 8, 2025.

This work was supported in part by the National Science and Technology Major Project (No. 2022ZD0116900), the National Natural Science Foundation of China (No. 52277118), and the Natural Science Foundation of Tianjin (No. 22JCZDJC00660).

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

Y. Zhou, Y. Peng, and L. Ge (corresponding author) are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: zhouyuan@tju.edu.cn; 202234073@tju.edu.cn; legendlj99@tju.edu.cn).

L. Hou is with the School of Intelligent Manufacturing Engineering and Future Technologies, Fuyao University of Science and Technology, Fuzhou 350001, China (e-mail: luyang.hou@hotmail.com).

Y. Wang and H. Niu are with the School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (e-mail: wangying01@mail.lzjtu.cn; nhx56055@mail.lzjtu.cn).

DOI: 10.35833/MPCE.2024.000643



to learn the operational environment of VVC. Reference [10] integrates graph neural networks into the DRL to model the VVC operation.

The aforementioned VVC methods focus on traditional discrete voltage regulation devices that operate on slower timescales, which may not be sufficient for the rapid and continuous control required by devices such as PV inverters and static var compensators (SVCs). Model-based optimization methods for the control of continuous voltage regulation devices can be categorized into centralized, distributed, and local control methods. Centralized control methods include a three-stage VVC method presented in [11], where the inverter supports both centralized reactive power scheduling and local reactive power droop control. Reference [12] introduces a reliability-constrained VVC method for PV inverters in distribution networks under uncertainty. Distributed control methods can be implemented through consensus-based methods [13] and distributed optimization methods [14], [15]. Local control methods involve dividing large distribution networks into smaller zones for automatic voltage control [16]-[18]. In the realm of DRL-based VVC methods designed for the rapid and continuous control of continuous voltage regulation devices, [19] proposes a voltage control method utilizing deep deterministic policy gradient (DDPG) to adjust the reactive power of multiple PV inverters. Reference [20] employs a multi-agent constrained soft actor-critic (MACSAC) algorithm to coordinate the reactive power of PV inverters. Reference [21] focuses on optimizing the continuous voltage regulation devices within minutes using a reinforcement learning (RL) algorithm based on graph convolution. Reference [22] utilizes two critics to independently learn the dual objectives of inverter-based VVC.

The model-based optimization methods and DRL-based VVC methods discussed earlier are applicable to scenarios where the distribution network comprises either traditional discrete voltage regulation devices such as OLTCs and CBs or exclusively continuous voltage regulation devices such as PV inverters. However, real-world distribution networks often contain a mix of both types of devices. Traditional discrete voltage regulation devices offer a wide adjustment range but operate at slower adjustment frequencies, making them unable to respond quickly to sudden voltage fluctuations. In contrast, continuous voltage regulation devices can make high-frequency adjustments but have a limited adjustment range. Therefore, it is crucial to consider the characteristics of both discrete and continuous voltage regulation devices when developing VVC methods for real-world distribution networks.

Several methods have been proposed to coordinate both discrete and continuous voltage regulation devices in distribution networks. References [23]-[25] propose model-based optimization methods to collaboratively optimize traditional voltage regulation devices and smart inverters operating on different timescales. Additionally, DRL-based VVC methods offer solutions for multi-device VVC. Reference [26] employs deep-Q network (DQN) to address optimization issues involving traditional voltage regulation devices. However, it relies on an accurate model for the optimization of continu-

ous voltage regulation devices. Reference [27] proposes a two-stage DRL method for VVC, where the first stage transforms the optimization problem of discrete voltage regulation devices into a model-based mixed-integer second-order cone program, and the second stage employs DRL to manage the actions of PV inverters. Reference [28] proposes a two-stage optimization method that controls discrete voltage regulation devices using historical data while employing DRL agents for real-time control of smart inverters. However, this method may be adversely affected by discrepancies between historical and real-time data, particularly concerning factors like weather conditions and load profiles. Reference [29] utilizes two separate DRL agents to control devices operating on different timescales. Reference [30] proposes a two-timescale coordination strategy that leverages electric springs for VVC; however, achieving an optimal solution requires an accurate circuit model.

In the realm of DRL-based VVC methods for distribution networks, algorithms such as DQN [31] and twin delayed deep deterministic policy gradient (TD3) [32] face limitations in their action spaces, being restricted to either discrete or continuous actions. This restricts their effectiveness in real-world scenarios where both types of control variables co-exist. Furthermore, when different algorithms or agents are employed to manage the two categories of devices independently, the interactions between their control actions are often overlooked. Consequently, there is a research gap in exploring RL algorithms with a hybrid action space that can simultaneously accommodate both discrete and continuous voltage regulation devices while considering the interactions between these devices within a unified decision-making framework.

To address this gap, this paper proposes a two-timescale VVC method based on RL with hybrid action space for distribution networks, termed the hybrid action representation twin delayed deep deterministic policy gradient (HAR-TD3) method. This method jointly controls traditional discrete voltage regulation devices operating on a slower timescale and smart continuous voltage regulation devices operating on a faster timescale. Specifically, to enable collaboration between the disparate action spaces of discrete and continuous voltage regulation devices, we introduce a variational autoencoder (VAE) [33] based hybrid action reconstruction network. During the reconstruction process, the encoder and decoder networks carefully account for the correlations between the two types of devices and the real-time status of the distribution network, facilitating coordinated control across multiple devices.

The contributions of this paper are summarized as follows.

- 1) A two-timescale VVC method based on RL with hybrid action space for distribution networks, termed the HAR-TD3 method, is proposed, which enables the simultaneous control of multiple devices in a model-free pattern.

- 2) The action space of the TD3 algorithm is improved. Specifically, the TD3 algorithm features two action spaces tailored for discrete and continuous voltage regulation devices within the distribution network. The actions of OLTCs and CBs are performed on a slower timescale, while the ac-

tions of PV inverters and SVCs are performed on a faster timescale.

3) A VAE-based hybrid action reconstruction network is proposed, which embeds discrete and continuous actions into the latent representation space. This network studies potential correlations between the actions of multiple types of devices in a distributed network.

The remainder of this paper is organized as follows. Section II introduces the framework of the proposed HAR-TD3 method. The Markov decision process (MDP) formulation of VVC problem is formulated in Section III. Section IV introduces the TD3 algorithm and the VAE-based hybrid action reconstruction network to solve the multi-device VVC problem. Section V validates the performance of proposed VVC method. Section VI concludes the whole paper.

## II. FRAMEWORK OF PROPOSED HAR-TD3 METHOD

The proposed HAR-TD3 method effectively oversees both discrete and continuous voltage regulation devices to ensure the voltages within the desired range. The framework for the proposed HAR-TD3 method is illustrated in Fig. 1.

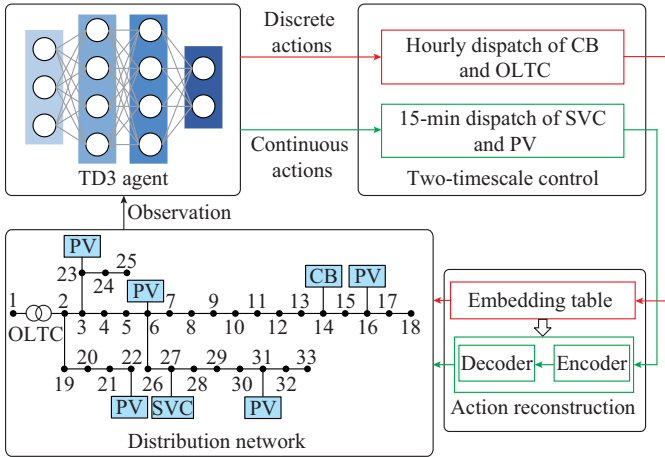


Fig. 1. Framework of proposed HAR-TD3 method.

The distribution network includes two types of devices: traditional discrete voltage regulation devices and smart continuous voltage regulation devices. The key challenge lies in controlling these devices, each with different action space, within a unified framework. To address this, we proposed the HAR-TD3 method. To accommodate the distinct control requirements of devices operating on different timescales, the proposed HAR-TD3 method divides a day into two timescales. The slow timescale operates with a control interval of 1 hour, with its time index represented as  $T = \{1, 2, \dots, N_T\}$ , where  $N_T = 24$ . The fast timescale operates with a control interval of 15 min, and the time index value is represented as  $t = \{1, 2, \dots, N_t\}$ , where  $N_t = 96$ . In this setup, each slow timescale period is divided into  $m$  fast timescales. Discrete voltage regulation devices such as CBs and OLTC operate on the slow timescale, while continuous voltage regulation devices such as PV inverters and SVCs operate on the fast timescale. The proposed HAR-TD3 method introduces a TD3 algorithm and a hybrid action reconstruction network to coor-

dinate discrete and continuous actions. The TD3 agent continuously receives real-time status observations from the distribution network. It updates the actions of discrete voltage regulation devices on the slow timescale and adjusts the actions of continuous voltage regulation devices on the fast timescale. Subsequently, the hybrid action reconstruction network builds an embedding table to represent discrete actions, and employs a VAE to map discrete actions' embedding and continuous actions' vectors into a latent representation space. The network then reconstructs the actions to their original dimensions and applies them to the distribution network.

## III. MDP FORMULATION OF VVC PROBLEM

RL models the interaction with the environment as an MDP, which is defined by a five-tuple  $(S, A, p, r, \gamma)$ .  $S$  is the state space, and  $s \in S$  is the state.  $A$  is the action space, and  $a \in A$  is the selected action.  $p$  is the state transition probability distribution.  $r$  is the immediate reward obtained after taking an action.  $\gamma$  is the discount factor that balances immediate and future rewards. In an MDP, an agent interacts with the environment over a sequence of time steps. At each time step, the agent observes the current state, selects an action based on a policy, transitions to a new state, and receives an immediate reward. The objective is to learn a policy to maximize cumulative rewards:

$$\mu^*(a_t | s_t) = \arg \max_{\mu} J(\mu) \quad (1)$$

$$J(\mu) = E \left[ \sum_{t=0}^{N_t} \gamma^t r_t \right] \quad (2)$$

where  $a_t$  and  $s_t$  represent the action and state at time step  $t$ , respectively;  $\mu$  represents the policy, which is a stochastic distribution of action  $a_t$  under state  $s_t$ , i.e.,  $a_t \sim \mu(\cdot | s_t)$ ;  $N_t$  is the length of episode;  $r_t$  is the immediate reward at time step  $t$ ;  $\gamma \in [0, 1]$ ; and  $E[\cdot]$  is the mathematical expectation.

In the VVC problem, both discrete and continuous voltage regulation devices must be coordinated to minimize long-term average voltage deviations and power losses. This coordination is achieved by leveraging real-time information collected from sensors in the distribution network, which serves as the foundation for decision-making. In this context, the power flow in the distribution network represents the environment, while the various controllable devices serve as control variables in the decision-making process of the agent. Consequently, the multi-device VVC problem can be formulated as an MDP, where the episodes, hybrid actions, states, and rewards are defined as follows.

### 1) Episode

In the MDP framework, an episode is defined as a single day, which is divided into steps of 15 min, resulting in a total of 96 steps per episode. Continuous voltage regulation devices operate with an action cycle of 15 min, meaning that they take an action every step. In contrast, discrete voltage regulation devices have an action cycle of 1 hour, resulting in an action being taken every four steps.

## 2) Hybrid actions

In the distribution network, the numbers of CBs, PV inverters, and SVCs in the distribution network are denoted as  $N_B$ ,  $N_{PV}$ , and  $N_{SVC}$ , respectively. The number of adjustable tap positions for OLTC is  $n^{OLTC}$ , and the number of stages of the  $i^{th}$  CB is  $n^{CB_i}$ . The vector  $\mathbf{a}_T^d$ , with length  $L_d$ , represents the discrete action information of discrete voltage regulation devices at time step  $T$ , where  $L_d = n^{OLTC} n^{CB_1} n^{CB_2} \dots n^{CB_{N_B}}$ . When interacting with the distribution network, the vector  $\mathbf{a}_T^d$  maps to specific discrete actions of discrete voltage regulation devices, i. e.,  $\mathbf{a}_T^d \rightarrow [a_T^{TAP}, a_{1,T}^{CB}, a_{2,T}^{CB}, \dots, a_{N_B,T}^{CB}]^T$ .  $a_T^{TAP}$  is the tap position of the OLTC, taking values from  $\{t_p^{\min}, \dots, -1, 0, 1, \dots, t_p^{\max}\}$ , where  $t_p^{\max}$  and  $t_p^{\min}$  are the maximum and minimum values of tap position, respectively;  $a_{i,T}^{CB}$  is the action of the  $i^{th}$  CB, taking values from  $\{0, 1, \dots, n^{CB_i}\}$ . The action of continuous voltage regulation devices at time step  $t$  is defined as  $\mathbf{a}_t^c = [a_{1,t}^{PV}, a_{2,t}^{PV}, \dots, a_{N_{PV},t}^{PV}, a_{1,t}^{SVC}, a_{2,t}^{SVC}, \dots, a_{N_{SVC},t}^{SVC}]^T$ , where  $a_{i,t}^{PV}$  and  $a_{i,t}^{SVC}$  are the actions of PV inverter and SVC, respectively, with values ranging continuously between  $-1$  to  $1$ . The length of vector  $\mathbf{a}_t^c$  is  $L_c = N_{PV} + N_{SVC}$ .

## 3) States

The state of MDP at step  $t$  is defined as a vector  $\mathbf{s}_t = [V_t, \mathbf{P}_t^{PV}, \mathbf{Q}_t^{SVC}, k_t, load_t, t]^T$ , where  $V_t$  is the vector of voltage magnitudes of all buses;  $\mathbf{P}_t^{PV} = [P_{1,t}^{PV}, P_{2,t}^{PV}, \dots, P_{N_{PV},t}^{PV}]^T$ , and  $P_{i,t}^{PV}$  is the active power output of the  $i^{th}$  PV inverter;  $\mathbf{Q}_t^{SVC} = [Q_{1,t}^{SVC}, Q_{2,t}^{SVC}, \dots, Q_{N_{SVC},t}^{SVC}]^T$ , and  $Q_{i,t}^{SVC}$  is the reactive power output of the  $i^{th}$  SVC;  $k_t$  is the index value of the discrete action taken; and  $load_t$  is the load power factor.

## 4) Rewards

$r_t$  is the reward at time step  $t$ . The reward encompasses two components, as follows:

$$r_t = -P_t^{\text{loss}} - pV_t^{\text{loss}} \quad (3)$$

$$P_t^{\text{loss}} = \sum_{ij \in \rho} P_{ij,t}^{\text{loss}} \quad (4)$$

$$V_t^{\text{loss}} = \sum_{i \in \mathcal{N}} \left( [V - V_{i,t}]_+ + [V_{i,t} - \bar{V}]_+ \right) \quad (5)$$

where  $P_t^{\text{loss}}$  is the active power loss at time step  $t$ ;  $ij$  is the branch from bus  $i$  to bus  $j$ ;  $P_{ij,t}^{\text{loss}}$  is the active power loss of branch  $ij$  at time step  $t$ ;  $\rho$  is the set of branches in the distribution network;  $p$  is a penalty factor for exceeding the voltage limit;  $V_t^{\text{loss}}$  is the voltage violation value at time step  $t$ ;  $\mathcal{N}$  is the set of buses in the distribution network;  $V$  and  $\bar{V}$  are the lower and upper voltage limits, respectively;  $V_{i,t}$  is the voltage at bus  $i$  at time step  $t$ ; and  $[x]_+ = \max(0, x)$ .

## IV. TD3 ALGORITHM AND VAE-BASED HYBRID ACTION RECONSTRUCTION NETWORK

### A. TD3 Algorithm

The TD3 algorithm is an enhancement of the DDPG algorithm that effectively mitigates issues such as  $Q$ -value overestimation and poor update [32], [34]. In the actor-critic

framework, the TD3 agent is composed of the critic network  $Q(\mathbf{s}, \mathbf{a} | \theta^Q)$  with parameter  $\theta^Q$ , and an actor network  $\mu(\mathbf{s} | \theta^\mu)$  with parameter  $\theta^\mu$ . Each actor and critic network is paired with a corresponding target network, denoted as  $Q'(\mathbf{s}, \hat{\mathbf{a}} | \theta^{Q'})$  for the critic and  $\mu'(\mathbf{s} | \theta^{\mu'})$  for the actor. The actor network continuously selects deterministic actions tailored to the input state, while the critic network evaluates these actions and provides gradient feedback to the actor network. These gradients help direct the actor network towards selecting progressively optimal actions.

TD3 algorithm has the following three characteristics.

Firstly, TD3 algorithm establishes two critic networks:  $Q_1(\mathbf{s}, \mathbf{a} | \theta^{Q_1})$  and  $Q_2(\mathbf{s}, \mathbf{a} | \theta^{Q_2})$ . Each of these critic networks has a corresponding target network, namely  $Q'_1(\mathbf{s}, \hat{\mathbf{a}} | \theta^{Q'_1})$  and  $Q'_2(\mathbf{s}, \hat{\mathbf{a}} | \theta^{Q'_2})$ . Two distinct action values of the next state are calculated by:

$$\begin{cases} Q'_1(\mathbf{s}_{t+1}, \hat{\mathbf{a}}_{t+1} | \theta^{Q'_1}) = Q'_1(\mathbf{s}_{t+1}, \mu'(\mathbf{s}_{t+1} | \theta^{\mu'}) | \theta^{Q'_1}) \\ Q'_2(\mathbf{s}_{t+1}, \hat{\mathbf{a}}_{t+1} | \theta^{Q'_2}) = Q'_2(\mathbf{s}_{t+1}, \mu'(\mathbf{s}_{t+1} | \theta^{\mu'}) | \theta^{Q'_2}) \end{cases} \quad (6)$$

To tackle the problem of  $Q$ -value overestimation in calculations, the TD3 algorithm computes the target  $Q$ -value by selecting the smaller value from the output of the two target critic networks. This target  $Q$ -value is crucial for calculating the loss functions for two critic networks, as shown in (7) and (8).

$$y_t = r_t + \gamma \min_{i=1,2} Q'_i(\mathbf{s}_{t+1}, \hat{\mathbf{a}}_{t+1} | \theta^{Q'_i}) \quad (7)$$

$$L(\theta^{Q_i}) = E \left[ \left( y_t - Q_i(\mathbf{s}_t, \mathbf{a}_t | \theta^{Q_i}) \right) \left( \mathbf{a}_t = \mu(\mathbf{s}_t | \theta^\mu) \right)^2 \right] \quad (8)$$

where  $y_t$  is the target  $Q$ -value at the specific state  $\mathbf{s}_t$  and action  $\mathbf{a}_t$ ; and  $L(\theta^{Q_i})$  is the loss function.

The critic network updates its parameters using gradient descent, as shown in (9) and (10).

$$\nabla_{\theta^{Q_i}} L(\theta^{Q_i}) = E \left[ \left( y_t - Q_i(\mathbf{s}_t, \mathbf{a}_t | \theta^{Q_i}) \right) \nabla_{\theta^{Q_i}} Q_i(\mathbf{s}_t, \mathbf{a}_t | \theta^{Q_i}) \right] \Big|_{i=1,2} \quad (9)$$

$$\theta^{Q_i} \leftarrow \theta^{Q_i} - \alpha \nabla_{\theta^{Q_i}} L(\theta^{Q_i}) \quad (10)$$

where  $\nabla_{\theta^{Q_i}} L(\theta^{Q_i})$  is the gradient of the loss function with respect to  $\theta^{Q_i}$ ;  $\nabla_{\theta^{Q_i}} Q_i(\mathbf{s}_t, \mathbf{a}_t | \theta^{Q_i})$  is the gradient of  $Q$ -value with respect to  $\theta^{Q_i}$ ; and  $\alpha$  is the learning rate of the critic network.

The actor network updates its parameter  $\theta^\mu$  using gradient ascent to maximize the expected cumulative reward. The optimization process adjusts the parameters to boost the  $Q$ -value as follows:

$$\begin{aligned} \nabla_{\theta^\mu} J(\theta^\mu) &= E \left[ \nabla_{\mathbf{a}} Q_1(\mathbf{s}, \mathbf{a} | \theta^{Q_1}) \Big| \mathbf{s} = \mathbf{s}_t, \right. \\ &\quad \left. \mathbf{a} = \mu(\mathbf{s}_t | \theta^\mu) \nabla_{\theta^\mu} \mu(\mathbf{s}_t | \theta^\mu) \Big| \mathbf{s} = \mathbf{s}_t \right] \end{aligned} \quad (11)$$

$$\theta^\mu \leftarrow \theta^\mu + \beta \nabla_{\theta^\mu} J(\theta^\mu) \quad (12)$$

where  $\nabla_{\theta^\mu} J(\theta^\mu)$  is the gradient of the target function  $J(\theta^\mu)$  with respect to  $\theta^\mu$ ;  $\nabla_{\mathbf{a}} Q_1(\mathbf{s}, \mathbf{a} | \theta^{Q_1})$  is the gradient of  $Q$ -value

with respect to  $\mathbf{a}$ ;  $\nabla_{\theta^\mu} \mu(\mathbf{s}|\theta^\mu)$  is the gradient of the policy  $\mu$  with respect to  $\theta^\mu$ ; and  $\beta$  is the learning rate of the actor network.

Secondly, to mitigate overfitting and reduce the variance of the temporal difference (TD) target, a clipped normal distribution noise  $\zeta$  is introduced to the target action, as expressed in (13).

$$\begin{cases} \hat{\mathbf{a}}_{t+1} \leftarrow \mu'(\mathbf{s}_{t+1}|\theta^{\mu'}) + \zeta \\ \zeta \sim \text{clip}(N(0, \sigma^2), -c, c) \end{cases} \quad (13)$$

where  $\zeta$  is sampled from a normal distribution  $N(0, \sigma^2)$  and clipped within the range  $[-c, c]$ ; and  $\text{clip}(x, -c, c)$  is the clipping function that limits  $x$  to a specified range  $[-c, c]$ , ensuring it stays within the interval by truncating values below  $-c$  or above  $c$ .

Thirdly, the actor network and target networks are updated less frequently than the critic networks, which can reduce the error accumulation during each update and stabilize the

learning process.

### B. Hybrid Action Space Improvements for TD3 Algorithm

Controlling both discrete and continuous actions simultaneously poses significant challenges for traditional RL agents. To address this challenge and effectively manage discrete and continuous voltage regulation devices in distribution networks, this paper focuses on the parametric action Markov decision process (PAMDP) [35]. PAMDP extends the standard MDP by introducing a hybrid action space that incorporates both discrete and continuous components.

Specifically, two enhancements have been introduced to the actor network and target actor network of the TD3 algorithm: hybrid action output and timescale constraint, as depicted in Fig. 2.  $\mathbf{a}_{\text{TD3},t}^d$  and  $\mathbf{a}_{\text{TD3},t}^c$  are the vectors representing discrete and continuous actions, respectively, output by the network at time step  $t$ . The actor network structure consists of multi-layer perceptron (MLP) function, rectified linear unit (ReLU) activation function, and tanh function.

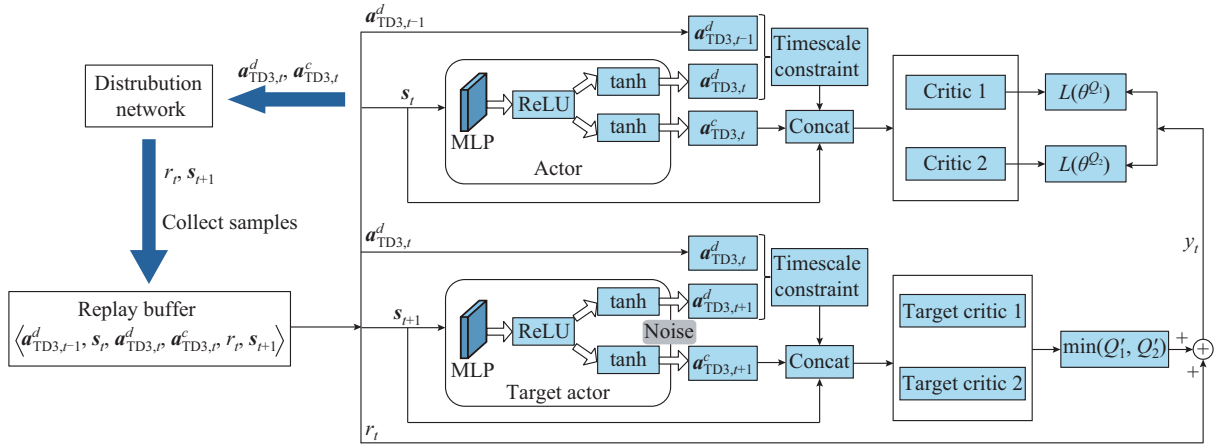


Fig. 2. Framework of TD3 algorithm.

#### 1) Hybrid Action Output

The actor network takes the state of the distribution network as input and outputs the corresponding actions, which should be executed in this state. The process is defined in (14) and (15).

$$\mathbf{a}_{\text{TD3},t}^d = \tanh\_d(\text{ReLU}(\text{MLP}(\mathbf{s}_t))) \quad (14)$$

$$\mathbf{a}_{\text{TD3},t}^c = \tanh\_c(\text{ReLU}(\text{MLP}(\mathbf{s}_t))) \quad (15)$$

where  $\text{MLP}(\cdot)$  is an MLP function consisting of multiple fully connected layers;  $\text{ReLU}(\cdot)$  is the ReLU activation function; and  $\tanh\_d(\cdot)$  and  $\tanh\_c(\cdot)$  are the hyperbolic tangent functions used to output discrete and continuous actions, respectively. The output dimensions of these actions are denoted as  $L_d$  for discrete actions and  $L_c$  for continuous actions.

#### 2) Timescale Constraint

To respect the timescale constraints associated with discrete actions of discrete voltage regulation devices, a specialized update mechanism is implemented. This mechanism ensures that discrete actions are updated less frequently compared with continuous actions, adhering to a slower timescale.

This constraint is described as:

$$\mathbf{a}_{\text{TD3},t}^d = \begin{cases} \mathbf{a}_{\text{TD3},t}^d & t \bmod m = 0 \\ \mathbf{a}_{\text{TD3},t-1}^d & \text{else} \end{cases} \quad (16)$$

where  $m$  is the number of continuous action executions that occur between two updates of discrete actions; and mod is the modulus operator, and when the result of  $t \bmod m$  is 0, it indicates that  $t$  is divisible by  $m$ . In this context,  $m=4$ .

#### C. Hybrid Action Reconstruction Network

In the collaborative reactive power optimization problem involving discrete and continuous voltage regulation devices, simply combining discrete and continuous actions without considering their inherent interdependence and heterogeneity can lead to suboptimal training outcomes. Inspired by recent advancements in RL and representation learning [36], the traditional hybrid action policy learning is transitioned to a continuous action policy learning in this paper. We introduce a latent representation space that captures the interactions and associations between discrete and continuous actions within the distribution network environment, aligning closely with

the dynamic characteristics of that environment. This subsection presents a network designed to construct a compact and interpretable latent representation space to effectively coordinate the actions of the devices. The structure of this network

is shown in Fig. 3, where  $k$  is the vector index in the embedding table;  $\delta_s$  is an output of the decoder to assist in predicting the system state; and  $z$  is a vector representing the latent representation variable of continuous actions.

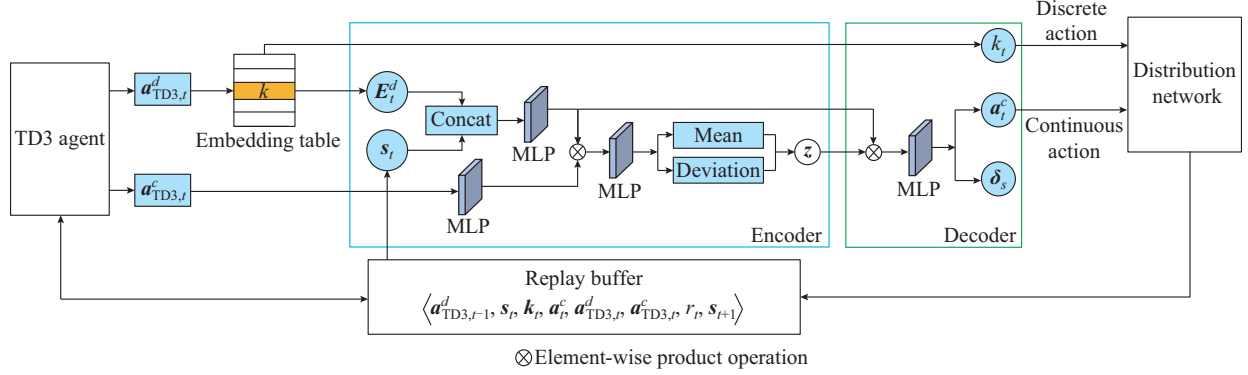


Fig. 3. Structure of hybrid action reconstruction network.

A learnable embedding table  $E \in \mathbb{R}^{K \times L_d}$  is established to represent the discrete actions, where  $K$  is the total number of discrete actions. Each row in the table corresponds to a discrete action and is represented by a continuous vector. The discrete actions are encoded by the embedding table as described in (17) and (18). Equation (17) calculates the L2-norm between the discrete action vector  $a_{TD3,t}^d$  and each row of the embedding table  $E(k)$ . The discrete action index  $k_t$  of the row with the smallest L2-norm is selected. Equation (18) derives the discrete action embedding vector corresponding to  $k_t$ , denoted as  $E_t^d$ , which serves as the encoding of action vector  $a_{TD3,t}^d$ .

$$k_t = \arg \min_{k \in K} \|E(k) - a_{TD3,t}^d\|_2 \quad (17)$$

$$E_t^d = E(k_t) \quad (18)$$

A VAE [33] is employed to construct the latent representation space for the continuous parameters. The VAE consists of an encoder  $q_\phi$ , parameterized by  $\phi$ , and a decoder  $p_\phi$ , parameterized by  $\phi$ . The encoder maps the continuous action vector  $a_{TD3,t}^c$ , the discrete action embedding vector  $E_t^d$ , and the state  $s_t$  into a latent representation variable  $z \in \mathbb{R}^l$ , as shown in (19). More concretely, the encoder outputs the mean  $\mu$  and standard deviation  $\sigma$  of a Gaussian distribution  $N(\mu, \sigma)$ , from which the latent representation space  $z$  is sampled.

$$z \sim q_\phi(\cdot | a_{TD3,t}^c, s_t, E_t^d) \quad (19)$$

The decoder then reconstructs continuous actions based on this latent representation space  $z$ , the discrete action embedding vector  $E_t^d$ , and the state  $s_t$ , as depicted in (20).

$$a_t^c = p_\phi(z, s_t, E_t^d) \quad (20)$$

Finally, the discrete action index  $k_t$  and the continuous action index  $a_t^c$  are applied to discrete and continuous voltage regulation devices in the distribution network, respectively.

The embedding table and VAE are trained by minimizing the loss function  $L_{VAE}$ , as follows:

$$L_{VAE} = E_{s_t, k_t, a_t^c \sim D, z \sim q_\phi} \left[ \|a_t^c - \tilde{a}_t^c\|_2^2 + D_{KL}(q_\phi(\cdot | a_t^c, s_t, E_t^d) \| N(0, 1)) \right] \quad (21)$$

where  $\tilde{a}_t^c$  is the output of the VAE during training;  $D$  is the replay buffer, which is used to sample data during training;  $E_{s_t, k_t, a_t^c \sim D, z \sim q_\phi}[\cdot]$  is the expectation operation, with  $s_t$ ,  $k_t$ , and  $a_t^c$  sampled from the replay buffer  $D$ , and  $z$  sampled from the encoder  $q_\phi$ ; the first term represents the L2-norm square reconstruction error; and the second term is the Kullback-Leibler (KL) divergence between the latent representation space  $z$  and the standard Gaussian prior  $N(0, 1)$ .

To enable the agent to evaluate the consequences of each device action in the distribution network more effectively, accurately predicting the next state of the system is crucial. This predictive capability allows the agent to make informed and optimized decisions about its actions. To enhance this capability, we incorporate an unsupervised learning loss based on the prediction of state dynamics. The construction of this loss function is detailed as:

$$L_S = E \left[ \|\delta_s - (s_{t+1} - s_t)\|_2^2 \right] \quad (22)$$

The overall loss function combines the VAE loss and the prediction loss of state dynamics, which is defined as:

$$L = L_{VAE} + \varepsilon L_S \quad (23)$$

where  $\varepsilon$  is a hyperparameter that balances the contribution of the prediction loss of state dynamics.

## V. CASE STUDY

### A. Simulation Settings

The proposed HAR-TD3 method is tested on modified IEEE 33-bus [37], IEEE 69-bus, and IEEE 123-bus distribution networks. In the modified IEEE 33-bus distribution network, an OLTC is placed between buses 1 and 2. The OLTC has 11 tap positions, each corresponding to a different turn ratio ranging from 0.95 to 1.05. A 4-stage CB is installed at bus 14 with a capacity of 400 kvar. Additionally, five PV

systems, each with a nameplate capacity of 200 kVA, are connected to the feeders at buses 6, 16, 22, 23, and 31. An SVC with a capacity of 200 kvar is installed at bus 27. In the modified IEEE 69-bus distribution network, an OLTC is placed between buses 1 and 2 with 11 tap positions, corresponding to turn ratios ranging from 0.95 to 1.05. Two 4-stage CBs, each with a capacity of 400 kvar, are located at buses 56 and 59. Furthermore, two SVCs, each with a capacity of 200 kvar, are installed at buses 62 and 64. Eight PV systems, each with a nameplate capacity of 200 kVA, are integrated into the feeders at buses 6, 16, 22, 23, 31, 34, 53, and 60. In the modified IEEE 123-bus distribution network, an OLTC is installed between buses 122 and 1, featuring 11 tap positions corresponding to turn ratios ranging from 0.95 to 1.05. Three 4-stage CBs, each with a capacity of 400 kvar, are installed at buses 64, 93, and 98. Four SVCs, each with a capacity of 400 kvar, are located at buses 52, 61, 96, and 97. Furthermore, 15 PV systems, each with a nameplate capacity of 600 kVA, are integrated into the feeders at buses 15, 20, 25, 26, 31, 42, 51, 60, 72, 80, 88, 93, 96, 100, and 109.

Electric loads and PV generation are sampled with a 15-min resolution. To account for voltage control uncertainties, random Gaussian noise with a standard deviation of 5% is introduced to both data, yielding multiple random scenarios. During the training process, each episode randomly selects a set of PV output and load data from the training dataset to create a variety of distribution network operating scenarios. The electric load is allocated across different buses according to the spatial load distribution specified in the original IEEE test feeder. All simulations are conducted using PyTorch, with training carried out on an NVIDIA GTX 1080Ti GPU and AMD Ryzen 2950X CPU to ensure efficient computational processing.

### B. Training Process and Parameter Settings

The proposed HAR-TD3 method incorporates multiple learnable networks, including the actor network, critic network, embedding table, encoder, and decoder. To enhance network performance and convergence, a multi-stage training approach is implemented, comprising three distinct stages. In the first stage, 1000 episodes of training samples are collected into the replay buffer. These samples are then used to pre-train the discrete action embedding table, encoder, and decoder over 5000 epochs. Following this pre-training phase, the parameters of the embedding table and VAE are frozen. In the second stage, the actor and critic networks are trained for 1000 episodes while new samples are continuously added to the replay buffer. During this phase, the VAE and embedding table remain unchanged to avoid destabilizing updates that could hinder policy learning. In the third stage, the embedding table and VAE are fine-tuned in conjunction with the ongoing training of the actor and critic networks until convergence is reached. This fine-tuning process enhances latent representations and optimizes the overall performance.

Pre-training the VAE components ensures that the encoder and decoder are well-initialized, which stabilizes the training

of the actor-critic network. Freezing the VAE during the initial actor-critic phase prevents destabilizing updates, thereby maintaining stable policy learning. In later stages, fine-tuning the VAE and embedding table allows the system to gradually enhance its latent representations and optimize overall performance, promoting more stable convergence and improved final outcomes. The parameter settings for training are detailed in Table I.

TABLE I  
PARAMETER SETTINGS FOR TRAINING

| Parameter                                                           | Value   |
|---------------------------------------------------------------------|---------|
| Number of replay buffers for agent                                  | 100000  |
| Number of replay buffers for hybrid action reconstruction network   | 1000000 |
| Number of mini-batch sizes for agent                                | 128     |
| Number of mini-batch sizes for hybrid action reconstruction network | 128     |
| Learning rate for critic network                                    | 0.0003  |
| Learning rate for actor network                                     | 0.0003  |
| Learning rate for hybrid action reconstruction network              | 0.0001  |
| Soft updating factor                                                | 0.001   |
| Discount factor                                                     | 0.99    |
| Standard deviation of noise                                         | 0.1     |
| Weight of prediction loss of state dynamics                         | 1       |

### C. Baseline Methods for Comparison

In the comparison study, the proposed HAR-TD3 method is evaluated against four state-of-the-art RL methods capable of managing hybrid discrete and continuous actions. The first baseline method is the proposed HAR-TD3 method without the hybrid action reconstruction network, termed “HAR-TD3 w/o VAE”. The second baseline method is parameterized action-space deep deterministic policy gradient (PA-DDPG) [38], designed for hybrid action spaces. The third baseline method is hierarchical hybrid Q-network (HHQN) [39], which consists of a high-level network responsible for coordinating discrete actions and a low-level network that learns the corresponding continuous parameters. The HHQN method employs centralized training for the high-level and low-level networks separately, followed by decentralized execution. The fourth baseline method is parameterized deep Q-network (PDQN) [40], which independently updates discrete action and continuous action policies by merging the structures of DQN and DDPG. All methods operate under a two-timescale setting and are trained on the same distribution network with identical data.

### D. Comparison of Voltage Control Performance

In this subsection, we evaluate the voltage control performance of each method using various metrics, including the 24-hour reward, 24-hour cumulative voltage violation (CVV) across all buses, and 24-hour total active power loss. The convergence curves for these metrics during the training with different methods are presented in Figs. 4-6. The solid line on the convergence curve is constructed by connecting points, where each point represents the average of the most

recent 100 episodes.

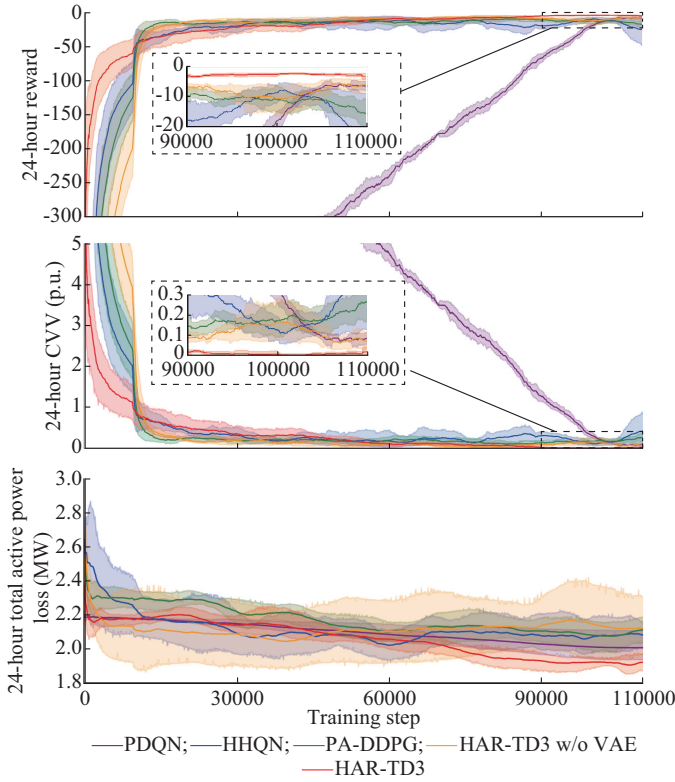


Fig. 4. Convergence curves for different metrics during training in modified IEEE 33-bus distribution network.

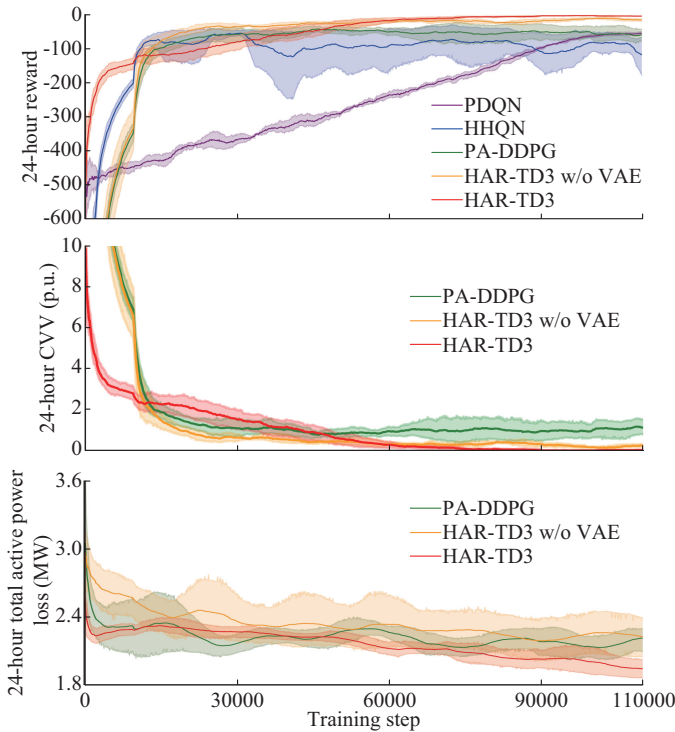


Fig. 5. Convergence curves for different metrics during training in modified IEEE 69-bus distribution network.

These averages are further computed across five random seeds to form the final solid line, while the shaded region

represents the range of variance based on the results from these random seeds. Higher values of the 24-hour reward curve signify better network training performance. To ensure stable operation, the voltages of all buses should ideally remain within the range of 0.95 to 1.05 p.u., meaning that the convergence value of the 24-hour CVV curve should approach zero. For the modified IEEE 33-bus distribution network shown in Fig. 4, the 24-hour reward curve indicates that while both the proposed HAR-TD3 method and baseline methods converge, the proposed HAR-TD3 method achieves a higher final convergence value compared with the baseline methods. In the modified IEEE 69-bus distribution network shown in Fig. 5, the convergence performance of the baseline methods deteriorates as the action and state space dimensions increase, while the proposed HAR-TD3 method maintains a significantly higher convergence value. In the modified IEEE 123-bus distribution network shown in Fig. 6, the baseline methods exhibit poor convergence performance. The HHQN method fails to converge, the PDQN method achieves a maximum reward of approximately  $-1500$ , and the PA-DDPG and HAR-TD3 w/o VAE methods reach their maximum rewards at around  $-500$ . In contrast, the proposed HAR-TD3 method successfully converges with a reward approaching 0.

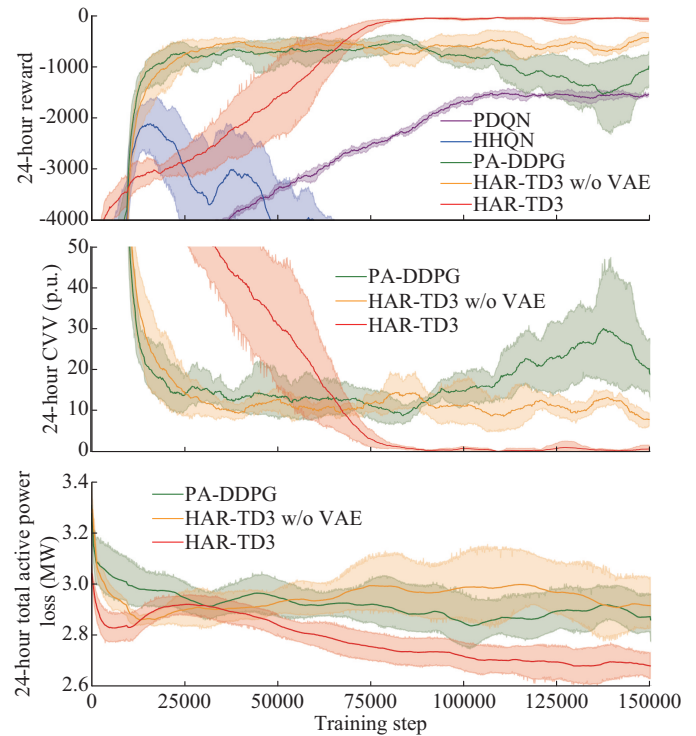


Fig. 6. Convergence curves for different metrics during training in modified IEEE 123-bus distribution network.

From the perspective of voltage control performance, an analysis of the 24-hour CVV curves in Figs. 4-6 indicates that the proposed HAR-TD3 method effectively reduces the CVV values to nearly zero. In contrast, the baseline methods struggle to coordinate multiple controllable devices in the distribution network, resulting in difficulties in maintaining voltages within the specified range. In the modified IEEE 33-

bus distribution network, while the baseline methods manage to control voltage within a relatively low violation range, their violation values are still higher than those achieved by the proposed HAR-TD3 method. Additionally, an examination of the total active power loss curves reveals that the baseline methods exhibit higher convergence values of total active power loss compared with the proposed HAR-TD3 method. In the modified IEEE 69-bus distribution network, the convergence of reward values indicates that, within the range of 60000 to 110000 training steps, the proposed HAR-TD3 method effectively reduces the CVV values to nearly zero while simultaneously minimizing the total active power loss. In the modified IEEE 123-bus distribution network, after 80000 training steps, the proposed HAR-TD3 method achieves significantly lower CVV and total active power loss compared with the baseline methods. The proposed HAR-TD3 method can effectively mitigate CVV values to nearly zero, whereas the baseline methods continue to exhibit relatively high CVV values.

As shown in Figs. 4-6, during the initial stages of training, the actor and critic networks in the proposed HAR-TD3 method require exploratory learning to adapt to the action reconstruction process of VAE, leading to slower convergence. However, as training progresses, the proposed HAR-TD3 method gradually outperforms the baseline methods in terms of final performance. This improvement can be attributed to the latent representation space constructed by the VAE network. Specifically, discrete actions are transformed into learnable vector representations through an embedding table. These vectors are then combined with continuous action vectors and fed into the VAE. After processing by the encoder of VAE, a unified latent representation space is generated. Throughout this process, the embedding table and encoder effectively capture the dynamic relationships between discrete and continuous actions, enabling the model to process both types of actions simultaneously within the latent representation space and enhancing decision-making accuracy.

Handling high-dimensional states and actions presents increasing complexity in extracting relevant state information and making accurate decisions. As illustrated in Fig. 6, integrating the VAE significantly enhances the model performance in such scenarios. This improvement is primarily due to the ability of the hybrid action reconstruction network to capture the dependencies between continuous and discrete actions while optimizing action selection through the generation of structured representations. By constructing a unified latent representation space, the hybrid action reconstruction network improves the representation of essential features in the decision-making process, effectively filtering out irrelevant or secondary information. This ability allows the model to efficiently extract useful information from high-dimensional data, thereby significantly improving decision-making accuracy.

The data presented in Table II indicate the mean value of the optimal convergence value of the 24-hour CCV and the corresponding total active power loss under five random seeds. It can be concluded that the proposed HAR-TD3 method outperforms the baseline methods in discrete and continuous multi-device VVC tasks.

TABLE II  
COMPARISON OF VVC PERFORMANCE OF DIFFERENT METHODS

| Distribution network | Method          | 24-hour CVV (p.u.) | Total active power loss (MW) |
|----------------------|-----------------|--------------------|------------------------------|
| IEEE 33-bus          | PDQN            | 0.0483744          | 2.01                         |
|                      | HHQN            | 0.0442815          | 2.05                         |
|                      | PA-DDPG         | 0.0436169          | 2.09                         |
|                      | HAR-TD3 w/o VAE | 0.0177633          | 2.13                         |
|                      | HAR-TD3         | 0.0002398          | 1.91                         |
| IEEE 69-bus          | PDQN            | 1.0391299          | 1.80                         |
|                      | HHQN            | 0.6560653          | 2.18                         |
|                      | PA-DDPG         | 0.5425647          | 2.17                         |
|                      | HAR-TD3 w/o VAE | 0.1173777          | 2.22                         |
|                      | HAR-TD3         | 0.0005848          | 1.98                         |
| IEEE 123-bus         | PDQN            | 27.8058360         | 2.86                         |
|                      | HHQN            | 39.2803702         | 2.87                         |
|                      | PA-DDPG         | 7.0804667          | 2.92                         |
|                      | HAR-TD3 w/o VAE | 4.9843449          | 2.93                         |
|                      | HAR-TD3         | 0.0014946          | 2.70                         |

When the power demand of an electric load exceeds the supply from a distributed PV system, voltage tends to decrease. During nighttime hours, when distributed PV systems are not generating power, inadequate adjustment of OLTC and reactive power compensation devices can result in excessively low voltage levels. Figures 7-9 illustrate the voltage conditions over a 24-hour period, where the red dashed line represents the upper or lower limit of the voltage. Prior to implementing control measures for each device in the distribution network, both continuous and discrete voltage regulation devices operate in an uncoordinated manner, leading to severe voltage violations. Figure 7 demonstrates that the proposed HAR-TD3 method effectively maintains voltage within the required range. However, during certain periods such as after the 19<sup>th</sup> hour, when PV generation decreases but a minor peak in electric load persists, the baseline methods struggle to accurately adjust the settings of OLTC and CB or to effectively increase the reactive power compensation of SVC. Consequently, the system experiences low voltage conditions, with several bus voltages dropping below 0.95 p.u..

As shown in Fig. 8, the proposed HAR-TD3 method exhibits exceptional voltage stability, ensuring that no bus voltage violations occur within the distribution network. In contrast, the baseline methods exhibit multiple instances of bus voltage violations under varying conditions. For instance, under the HHQN method, during the period from the 11<sup>th</sup> hour to the 15<sup>th</sup> hour, the combination of high PV power output and improper device operation results in numerous bus voltage readings exceeding the threshold of 1.05 p.u.. Similarly, during the peak consumption period from the 8<sup>th</sup> hour to the 20<sup>th</sup> hour, the PA-DDPG method results in a significant number of bus voltage readings falling below the lower threshold of 0.95 p.u.. Additionally, the HAR-TD3 w/o VAE method experiences instances where bus voltage readings fall below 0.95 p.u. between the 23<sup>th</sup> hour and midnight.

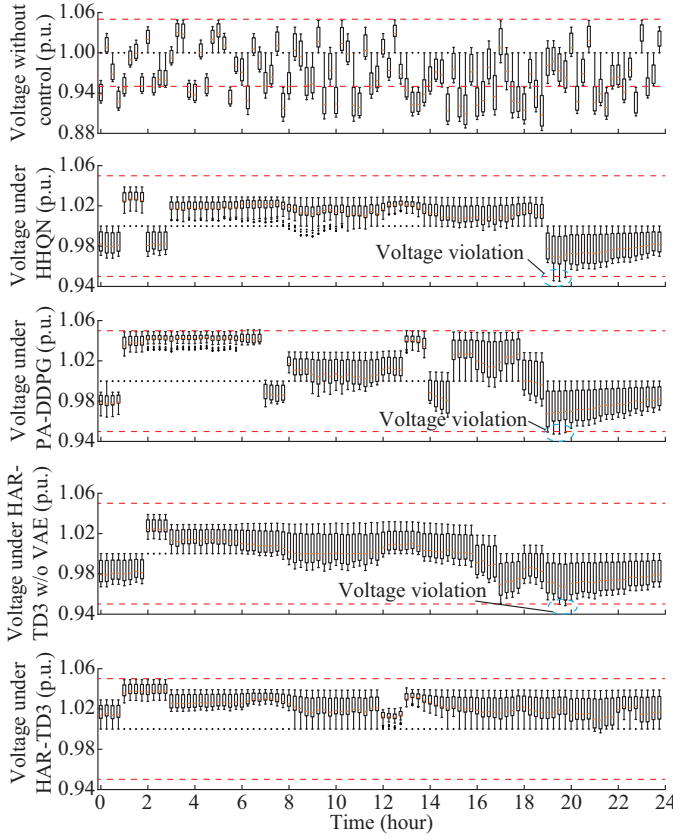


Fig. 7. Voltage of IEEE 33-bus distribution network controlled by different methods.

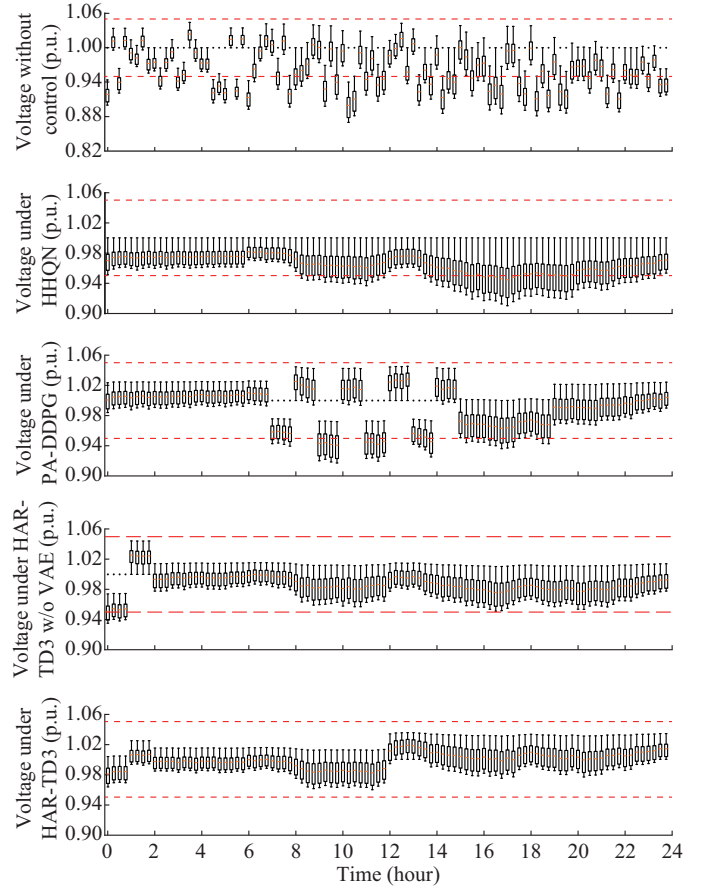


Fig. 9. Voltage of IEEE 123-bus distribution network controlled by different methods.

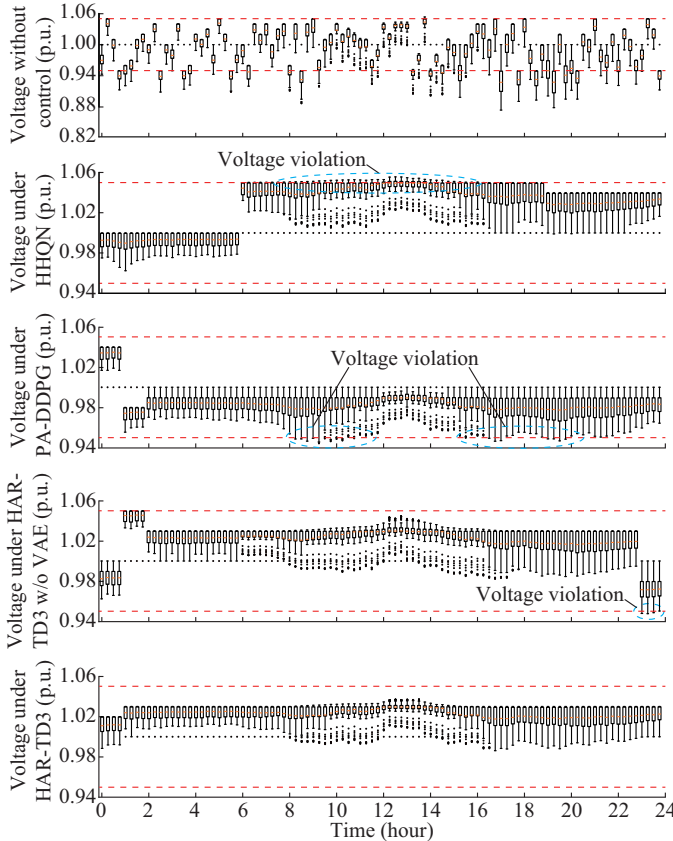


Fig. 8. Voltage of IEEE 69-bus distribution network controlled by different methods.

As shown in Fig. 9, the proposed HAR-TD3 method effectively manages device operations, maintaining bus voltage consistently within the specified range and avoiding the significant voltage violations observed under the baseline methods.

The comparison of voltage control performance across various methods, as presented in Figs. 4-9 and Table II, unequivocally showcases the superiority of the proposed HAR-TD3 method in terms of both convergence performance and voltage control efficacy. By incorporating the hybrid action space, the proposed HAR-TD3 method adeptly coordinates devices with disparate action spaces across two distinct timescales. Additionally, the introduction of the VAE-based hybrid action reconstruction network further enhances the overall voltage control performance. In this framework, each discrete action is meticulously mapped to a continuous vector within a customizable embedding table. Subsequently, the VAE is utilized to forge a latent representation space, grounded on the state and action vectors of the system. This innovative process adeptly captures the intricate dynamic correlations between discrete and continuous actions, thereby facilitating optimal decision-making for both types of devices operating on different timescales.

#### E. Discussion on State Prediction

Monitoring the changes in the state of the distribution net-

work is crucial for assessing the long-term implications of decision-making. To investigate the effect of the hyperparameter  $\varepsilon$  in (22) on voltage control performance, Fig. 10 depicts the daily trend of the total CVV in the IEEE 33-bus system. As Fig. 10 shows, when  $\varepsilon$  ranges between 0 and 2, the total CVV value decreases significantly and reaches its lowest point when  $\varepsilon$  is between 1 and 2. This result underscores the positive impact of incorporating the prediction loss of state dynamics on voltage control performance. During this range,  $\varepsilon$  is set at a moderate level, resulting in a significant improvement in the voltage control of the system. However, when  $\varepsilon$  exceeds 2, the total CVV value begins to rise sharply. This increase stems from the fact that an excessively high  $\varepsilon$  prompts the VVC method to focus excessively on predicting the future state of the distribution network, while neglecting the decision-making for the current state. This imbalance adversely affects voltage control performance. In summary, within a suitable range of  $\varepsilon$ , the prediction loss of state dynamics aids the VVC method in effectively balancing the short-term and long-term impact of current decisions, thereby further enhancing the overall voltage control performance.

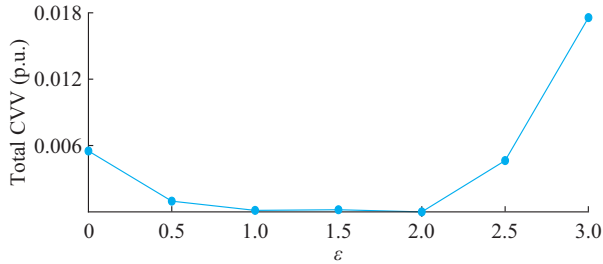


Fig. 10. Impact of hyperparameter  $\varepsilon$  on total CVV.

#### F. Comparison of Computation Time

Table III provides a comparison of the computation time required for different methods to execute a single episode, which comprises 96 decisions over a 24-hour period, across three different systems. The computation time of the DRL-based VVC methods primarily depends on the model size and the input data scale, often falling within the millisecond range. As shown in Table III, the computation time of the proposed HAR-TD3 method marginally increases compared with those of the baseline methods. This increase is primarily due to the higher model complexity. Specifically, in the IEEE 123-bus system, while the proposed HAR-TD3 method exhibits a slightly longer computation time, it achieves a significant reduction in voltage violations by several orders of magnitude compared with the baseline methods. Notably, the total computation time for all decisions throughout the day increases by only approximately 20 ms, which remains in an acceptable range. These findings underscore the computational efficiency of the proposed HAR-TD3 method. Despite its higher model complexity, the proposed HAR-TD3 method exhibits superior voltage control performance and reduced power loss without incurring a significant increase in computation time.

TABLE III  
COMPARISON OF COMPUTATION TIME FOR DIFFERENT METHODS

| Method          | Computation time per episode (ms) |                                  |                                   |
|-----------------|-----------------------------------|----------------------------------|-----------------------------------|
|                 | IEEE 33-bus distribution network  | IEEE 69-bus distribution network | IEEE 123-bus distribution network |
| PDQN            | 25.3                              | 27.4                             | 31.2                              |
| HHQN            | 22.2                              | 23.6                             | 25.7                              |
| PA-DDPG         | 17.4                              | 18.5                             | 21.3                              |
| HAR-TD3 w/o VAE | 17.4                              | 18.6                             | 21.5                              |
| HAR-TD3         | 42.8                              | 43.9                             | 48.6                              |

#### VI. CONCLUSION

This paper introduces a VVC method based on RL with hybrid action space, termed HAR-TD3 method. The proposed method simultaneously manages OLTC and CBs on a slower timescale, while controlling PV inverters and SVCs on a faster timescale in a model-free pattern. Specifically, an improved TD3 algorithm allows for the execution of both discrete and continuous actions across different timescales. To further optimize voltage control performance, we present a VAE-based hybrid action reconstruction network. This network maps discrete actions of discrete voltage regulation devices and continuous actions of continuous voltage regulation devices into a unified latent representation space, facilitating improved collaboration among devices. Case studies conducted on the modified IEEE 33-bus, 69-bus, and 123-bus distribution networks demonstrate that the proposed method outperforms existing DRL-based VVC methods in terms of convergence speed and voltage control effectiveness.

A promising direction for future research avenue involves extending the proposed method to develop a distributed VVC method, enabling coordinated and partitioned control of discrete and continuous voltage regulation devices.

#### REFERENCES

- [1] H. Ahmadi, J. R. Marti, and H. W. Dommel, "A framework for volt-var optimization in distribution systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1473-1483, May 2015.
- [2] N. Daratha, B. Das, and J. Sharma, "Coordination between OLTC and SVC for voltage regulation in unbalanced distribution system distributed generation," *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 289-299, Jan. 2014.
- [3] R. A. Shayani and M. A. G. de Oliveira, "Photovoltaic generation penetration limits in radial distribution systems," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1625-1631, Aug. 2011.
- [4] I. Roytelman, B. K. Wee, and R. L. Lugtu, "Volt/var control algorithm for modern distribution management system," *IEEE Transactions on Power Systems*, vol. 10, no. 3, pp. 1454-1460, Aug. 1995.
- [5] A. Padilha-Feltrin, D. A. Q. Rodezno, and J. R. S. Mantovani, "Volt-var multiobjective optimization to peak-load relief and energy efficiency in distribution networks," *IEEE Transactions on Power Delivery*, vol. 30, no. 2, pp. 618-626, Apr. 2015.
- [6] M. Manbachi, A. Sadu, H. Farhangi *et al.*, "Real-time co-simulation platform for smart grid volt-var optimization using IEC 61850," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 4, pp. 1392-1402, Aug. 2016.
- [7] D. Cao, W. Hu, J. Zhao *et al.*, "Reinforcement learning and its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, Dec. 2020.
- [8] H. Wu, Z. Xu, M. Wang *et al.*, "Full-model-free adaptive graph deep deterministic policy gradient model for multi-terminal soft open point

- voltage control in distribution systems,” *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 6, pp. 1893-1904, Nov. 2024.
- [9] Y. Gao and N. Yu, “Model-augmented safe reinforcement learning for volt-var control in power distribution networks,” *Applied Energy*, vol. 313, p. 118762, May 2022.
  - [10] X. Y. Lee, S. Sarkar, and Y. Wang, “A graph policy network approach for volt-var control in power distribution systems,” *Applied Energy*, vol. 323, p. 119530, Oct. 2022.
  - [11] C. Zhang, Y. Xu, Z. Dong *et al.*, “Three-stage robust inverter-based voltage/var control for distribution networks with high-level PV,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 782-793, Jan. 2019.
  - [12] Q. Chai, C. Zhang, Y. Xu *et al.*, “PV inverter reliability-constrained volt/var control of distribution networks,” *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1788-1800, Jul. 2021.
  - [13] J. Schiffer, T. Seel, J. Raisch *et al.*, “Voltage stability and reactive power sharing in inverter-based microgrids with consensus-based distributed voltage control,” *IEEE Transactions on Control Systems Technology*, vol. 24, no. 1, pp. 96-109, Jan. 2016.
  - [14] H. J. Liu, W. Shi, and H. Zhu, “Distributed voltage control in distribution networks: online and robust implementations,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6106-6117, Nov. 2018.
  - [15] T. Xu and W. Wu, “Accelerated ADMM-based fully distributed inverter-based volt/var control strategy for active distribution networks,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7532-7543, Dec. 2020.
  - [16] H. Zhu and H. Liu, “Fast local voltage control under limited reactive power: optimality and stability analysis,” *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3794-3803, Sept. 2016.
  - [17] W. Zheng, W. Wu, B. Zhang *et al.*, “A fully distributed reactive power optimization and control method for active distribution networks,” *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1021-1033, Mar. 2016.
  - [18] P. Jahangiri and D. C. Aliprantis, “Distributed volt/var control by PV inverters,” *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3429-3439, Aug. 2013.
  - [19] C. Li, C. Jin, and R. Sharma, “Coordination of PV smart inverters using deep reinforcement learning for grid voltage regulation,” in *Proceedings of 2019 18th IEEE International Conference on Machine Learning and Applications*, Boca Raton, USA, Dec. 2019, pp. 1930-1937.
  - [20] H. Liu and W. Wu, “Online multi-agent reinforcement learning for decentralized inverter-based volt-var control,” *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2980-2990, Jul. 2021.
  - [21] R. Yan, Q. Xing, and Y. Xu, “Multi-agent safe graph reinforcement learning for PV inverters-based real-time decentralized volt/var control in zoned distribution networks,” *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 299-311, Jan. 2024.
  - [22] Q. Liu, Y. Guo, L. Deng *et al.*, “Two-critic deep reinforcement learning for inverter-based volt-var control in active distribution networks,” *IEEE Transactions on Sustainable Energy*, vol. 15, no. 3, pp. 1768-1781, Jul. 2024.
  - [23] Y. Xu, Z. Y. Dong, R. Zhang *et al.*, “Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems,” *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4398-4408, Nov. 2017.
  - [24] C. Li, V. R. Disfani, H. V. Haghi *et al.*, “Optimal voltage regulation of unbalanced distribution networks with coordination of OLTC and PV generation,” in *Proceedings of 2019 IEEE PES General Meeting*, Atlanta, USA, Aug. 2019, pp. 1-5.
  - [25] R. R. Jha, A. Dubey, C. C. Liu *et al.*, “Bi-level volt-var optimization to coordinate smart inverters with voltage control devices,” *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1801-1813, May 2019.
  - [26] W. Wang, N. Yu, Y. Gao *et al.*, “Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems,” *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.
  - [27] X. Sun and J. Qiu, “Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method,” *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2903-2912, Jul. 2021.
  - [28] D. Hu, Z. Ye, Y. Gao *et al.*, “Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization,” *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4873-4886, Nov. 2022.
  - [29] F. Kabir, N. Yu, Y. Gao *et al.*, “Deep reinforcement learning-based two-timescale volt-var control with degradation-aware smart inverters in power distribution systems,” *Applied Energy*, vol. 335, p. 120629, Apr. 2023.
  - [30] J. Gong and K.-W. Lao, “Electric spring for two-timescale coordination control of voltage and reactive power in active distribution networks,” *IEEE Transactions on Power Delivery*, vol. 39, no. 3, pp. 1864-1876, Jun. 2024.
  - [31] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529-533, Feb. 2015.
  - [32] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 1587-1596.
  - [33] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada, Apr. 2014, pp. 1-14.
  - [34] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al.* (2015, Sept.). Continuous control with deep reinforcement learning. [Online]. Available: <http://arxiv.org/abs/1509.02971>
  - [35] W. Masson, P. Ranchod, and G. Konidaris, “Reinforcement learning with parameterized actions,” in *Proceedings of the 30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, Phoenix, USA, Feb. 2016, pp. 1934-1940.
  - [36] B. Li, H. Tang, Y. Zheng *et al.*, “Hyar: addressing discrete-continuous action reinforcement learning via hybrid action representation,” in *Proceedings of the 10th International Conference on Learning Representations*, Virtual Event, Apr. 2022, pp. 1-22.
  - [37] M. E. Baran and F. F. Wu, “Network reconfiguration in distribution systems for loss reduction and load balancing,” *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401-1407, Apr. 1989.
  - [38] M. J. Hausknecht and P. Stone, “Deep reinforcement learning in parameterized action space,” in *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016, pp. 1-12.
  - [39] H. Fu, H. Tang, J. Hao *et al.*, “Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, Aug. 2019, pp. 2329-2335.
  - [40] J. Xiong, Q. Wang, Z. Yang *et al.* (2018, Oct.). Parametrized deep Q-networks learning: reinforcement learning with discrete-continuous hybrid action space. [Online]. Available: <http://arxiv.org/abs/1810.06394>
- Yuan Zhou** received the B.Eng., M.Eng., and Ph.D. degrees in electronic engineering and communication engineering from Tianjin University, Tianjin, China, in 2006, 2008, and 2011, respectively. Since 2011, she has been a Faculty Member with the School of Electronic Information Engineering, Tianjin University, where she is currently a Professor. From 2013 to 2014, she was a Visiting Scholar with the School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, Australia. From 2016 to 2017, she was a Visiting Scholar with the Department of Electrical Engineering, Princeton University, Princeton, USA. Her research interests include computer vision, reinforcement learning, and active distribution system operation and control.
- Yizhou Peng** received the B.Eng. degree from Shandong University, Qingdao, China, in 2022, and is currently pursuing the M.Eng. degree in Tianjin University, Tianjin, China. Her research interests include reinforcement learning and active distribution system operation and control.
- Leijiao Ge** received Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2016. He is currently an Associate Professor in the School of Electrical and Information Engineering at Tianjin University. His main research interests include smart distribution network, cloud computing, and big data.
- Luyang Hou** received the Ph.D. degree in information system engineering from Concordia University, Montreal, Canada, in 2020. He is currently an Assistant Professor at the School of Intelligent Manufacturing Engineering and Future Technologies, Fuyao University of Science and Technology (FY-UST), Fuzhou, China. Before joining FYUST, he was a Research Associate Professor at Beijing University of Posts and Telecommunications, Beijing, China, and a Postdoctoral Research Fellow at The University of British Columbia, Vancouver, Canada. His research interests include operational optimization, mechanism design, and machine learning with applications to energy internet and intelligent manufacturing system. He also develops AI-eco-

conomic paradigm and applies it to integrated energy and information network in terms of smart factory scheduling, energy management, vehicle-grid interaction, electricity-carbon coordination, etc.

**Ying Wang** received the Ph.D. degree in power system and its automation from Southwest Jiaotong University, Sichuan, China, in 2016. He is currently a Full Professor with the School of Electrical Engineering and Automation, Lanzhou Jiaotong University, Lanzhou, China. His research interests in-

clude electrical contact of pantograph-catenary dynamic, power quality of electric traction system, and their applications in electrified railway industry.

**Hongxia Niu** received master's degree in traffic information engineering and control engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2003. She is currently an Associate Professor in the School of Automation and Electrical Engineering, Lanzhou Jiaotong University. Her research interests include computer vision and intelligent control.