Robust Voltage Control for Active Distribution Networks via Safe Deep Reinforcement Learning Against State Perturbations

Meng Tian, Xiaoxu Li, Ziyang Zhu, Zhengcheng Dong, Li Gong, and Jingang Lai

***Abstract*—With the prevalence of renewable distributed energy resources (DERs) such as photovoltaics (PVs), modern active distribution networks (ADNs) suffer from voltage deviation and power quality issues. However, traditional voltage control methods often face a trade-off between efficiency and effectiveness, and rarely ensure robust voltage safety under typical state perturbations in practical distribution grids. In this paper, a robust model-free voltage regulation approach is proposed which simultaneously takes security and robustness into account. In this context, the voltage control problem is formulated as a constrained Markov decision process (CMDP). A safety-augmented multi-agent deep deterministic policy gradient (MADDPG) algorithm is the trained to enable real-time collaborative optimization of ADNs, aiming to maintain nodal voltages within safe operational limits while minimizing total line losses. Moreover, a robust regulation loss is introduced to ensure reliable performance under various state perturbations in practical voltage controls. The proposed regulation algorithm effectively balance efficiency, safety, and robustness, and also demonstrates potential for generalizing these characteristics to other applications. Numerical studies validate the robustness of the proposed method under varying state perturbations on the IEEE test cases and the optimal integrated control performance when compared to other benchmarks.**

***Index Terms*—Active distribution network, robust voltage control, state perturbation, model-free, safe deep reinforcement learning.**

I. Introduction

In recent years, with environmental policies being implemented worldwide, the penetration of renewable distributed energy resources (DERs) in active distribution networks (ADNs), particularly photovoltaics (PVs), has been rapidly increasing [1]. However, PV generation is inherently uncertainty, intermittent, and unstable [2]–[3], as it is strongly influenced by external stochastic factors, often resulting in noticeable voltage fluctuations [4]. In addition, local mismatches between generation and demand can give rise issues such as reverse power flow [5] and power quality degradation [6], which further undermine voltage stability [7] and limit the system's ability to accommodate additional PV capacity [8]. Hence, effective voltage regulation is imperative for ensuring reliable ADN operation under high level of PV integration.

Typical voltage control consists of two main branches: droop control and stochastic programming of power flow. Droop control tunes PV inverters' reactive power injection or absorption via pre-designed local voltage-reactive power curves to regulate nodal voltages in ADN [9]. However, as a fully distributed method, it is susceptible to local measurement errors and difficult to obtain a global optimal solution [10]. Meanwhile, the manually designed control curves struggle to perform effectively under dynamic power system conditions [11]–[12]. Voltage control can be transformed into an optimal power flow problem with voltage as a constraint, which is often formulated as a nonlinear stochastic programming issue by the nature of line power flow equations. In existing literatures, the optimization can be solved either in a centralized manner, such as mixed-integer second order cone programming [13], genetic algorithms [14], and interior point methods [15], or by distributed approaches, such as alternating direction method of multipliers (ADMM) [16]–[17], distributed quasi-Newton's algorithms [18], and distributed heavy ball algorithms [19]. Notably, the performance of these model-based approaches towards voltage control is highly correlated with the precise physical model of ADNs [20]. However, due to the complex topology and uncertain parameters of real distribution

Received: February 1, 2025
Accepted: July 23, 2025
Published Online: January 1, 2026

Meng Tian and Zhengcheng Dong are with the School of Automation, Wuhan University of Technology, Wuhan 430070, China (e-mail: tm@whut.edu.cn; zcdong@whut.edu.cn).

Xiaoxu Li, Ziyang Zhu (corresponding author), and Li Gong are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: xylonlee@whu.edu.cn; ziyangzhu@whu.edu.cn; ligong@whu.edu.cn).

Jingang Lai is with the School of Artificial Intelligence and Automation, and also with the Key Laboratory of Image Processing and Intelligent Control, Education Ministry of China, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: kklai@hust. edu.cn).

DOI: 10.23919/PCMP.2024.000342

networks, accurately simulating the on-site system is nearly impossible, which in turn compromises the effectiveness of these algorithms. Moreover, model-based methods may encounter significant computational bottlenecks when accommodating large numbers of DERs [21]. As a result, their ability to meet the real-time control requirements of modern ADN may be compromised.

With the growing adoption of artificial intelligence technologies, advanced control algorithms, particularly those based on deep reinforcement learning (DRL), have been applied to a wide range of power system tasks [22]–[25], partially overcoming the limitations of traditional voltage control methods. DRL is a fully data-driven approach uses operational data interacting with the environment to pre-train decision models, which can achieve real-time on-the-fly management for fast-responding voltage regulation devices, such as PV inverters [26]–[28], static var compensators (SVCs) [29]–[30], etc. In addition, when the PV penetration in ADN is extremely high, the multi-agent reinforcement learning (MARL) methods adopt the divide-and-conquer philosophy, and can circumvent the dimensional explosion problem caused by centralized management and considerably reduce the computational burden of the controller. In order to enhance the collaboration of agents, MARL-based voltage control methods tend to utilize the algorithms with centralized training and decentralized execution (CTDE) framework, such as multi-agent deep Q network (MADQN) [31], multi-agent deep deterministic policy gradient (MADDPG) [32]–[35] and multi-agent soft actor-critic (MASAC) [36]–[37], to achieve a remarkable trade-off between communication cost and globally optimal decision-making. In general, DRL-based voltage control methods involve voltage constraints and energy loss as the objective function for optimization. However, in pursuing overall optimum, these methods do not always ensure that voltages remain within safe operating limits [38], which may pose risks to the safe operation of ADNs. To address this problem, safe deep reinforcement learning (SDRL) has been widely explored. Reference [39] formulates the optimal operation problem of distribution networks as a constrained Markov decision process (CMDP), utilizing the constrained policy optimization (CPO) algorithm to achieve minimal power cost while keeping voltages under regulation by coordinating conventional volt-var control (VVC) devices. An online multi-agent constrained soft actor-critic (MACSAC) with decentralized control framework is developed to fulfill VVC against physical model uncertainty [40], whereas in [41], a model-free constrained soft actor-critic (CSAC) algorithm is proposed for VVC, which formulates the voltage violations as constraints instead of rewards. Reference [42] presents a DRL algorithm associated with a safety module in the medium resolution of three stages VVC to schedule PV inventers and energy storage systems (ESSs), which eliminates voltage deviations throughout the online training phase. However, the aforementioned methods have slightly lower ceilings on voltage stability and power loss attenuation comparing to the conventional model-based ones, and meanwhile they require perturbation-free observational states to achieve decent performance.

Uncertainty problems are widely pervasive throughout power systems, such as disturbances in measured states [43]–[44], nebulous grid parameters and topologies [45]–[47], and uncertainties of PVs and loads [48]–[50]. Such indeterminacy significantly degrades the reliability and robustness of voltage control, with state disturbances being the most prominent. The state perturbation generally refers to misperception of present voltage or power flow profile across the whole ADN, typically caused by instrument errors [51], which significantly affects the decision-making ability of voltage controllers to coordinate active and reactive devices. Addressing the impact of these uncertainties on distribution network optimization strategies, enhancing algorithm robustness is increasingly becoming a focal issue in the field of voltage control in distribution networks recently [52]. In [53], a robust optimization method is also introduced to obtain Pareto solutions under uncertainties, and a multi-objective adaptive VVC is proposed to address voltage issues in ADNs with high PV penetration. Reference [54] introduces a two-stage approach with distributed optimization for classical voltage control devices and robust optimization for PV inverters, employing a non-centralized VVC algorithm. A universal distributionally robust safety filter for VVC is developed to use robust optimization with chance constraints to ensure near-optimal solutions and constraint satisfaction [55]. Considering bad data in measurement, reference [56] proposes a robust VVC approach by combining revised ADMM and local feedback control. Nevertheless, the aforementioned methods primarily leverages the principles of robust optimization. Although robust optimization is widely applied to address robust voltage control issues in distribution networks considering uncertainties, it still faces the issue of overly conservative optimization strategies due to its focus on the "worst-case" scenario. This highlights a new direction for subsequent research. Among various uncertainties, this paper primarily considers the impact of state perturbations on distribution network voltage optimization.

This paper proposes a robust multi-agent safe reinforcement learning method for voltage regulation of ADNs under state perturbations, referred to as robust safe MADDPG (RS-MADDPG). The main contributions of this paper are summarized as follows:

1) To ensure the operational safety of ADNs, the voltage control problem is modeled as a CMDP, and a safe MADDPG (S-MADDPG) algorithm is developed by incorporating techniques such as independent cost networks and parameter sharing during training. Significant improvements are observed in the voltage safety of ADNs compared with vanilla DRL-based methods such as MADDPG, highlighting the effectiveness and superiority of the proposed method on

addressing voltage control challenges in dynamic distribution network environments.

2) In view of ubiquitous state perturbations encountered in real-world ADNs, a novel component, namely, robust regulation loss, is introduced for SDRL-based voltage control. This mechanism is designed to balance safety and robustness in distribution network optimization, thereby enhancing practical applicability. Moreover, this innovation is not only compatible with the proposed S-MADDPG but also has the potential to generalize to other MARL-based algorithms for robust voltage regulation.

The rest of this paper is organized as follows. Section II describes the formulation of voltage control problem, while Section III introduces the proposed RS-MADDPG method. Section IV reports the outcomes of numerical study, and finally, Section V presents the overall conclusion.

II. Problem Formulation

In ADNs, each bus is considered as a node, whose set is denoted as $\mathcal{N}$. Among them, node 0 is the reference node fed by the transmission network, whose voltage magnitude and phase angle are defined as 1 p.u. and 0, respectively. The set of branch lines connected to node i is indicated as $\mathcal{B}^i (i \in \mathcal{N})$. Let $\mathcal{G} \in \mathcal{N}$ denote the set of nodes equipped with controllable PV inverters. The objective of voltage control is to minimize the deviation of the nodal voltage V^i from the standard reference level $V_{\text{ref}} = 1.00\,\text{p.u.}$, and the energy loss of entire lines, which is formulated as:

$$\min_{P_{\text{cur}}, Q_{\text{PV}}} \sum_{t=1}^{T} \left[\frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \left| V^i(t) - V_{\text{ref}} \right| + \frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \left(\alpha_q \left| Q_{\text{PV}}^j(t) \right| + \alpha_p P_{\text{cur}}^j(t) \right) \right] \tag{1}$$

where P_{cur} and Q_{PV} denote the active power curtailment of PVs and the reactive injection of inverters, respectively; $\frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \left| Q_{\text{PV}}^j(t) \right|$ is employed to minimize the reactive power generation while reducing line loss to a certain extent [24], and the curtailed power cost $\frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} P_{\text{cur}}^j(t)$ is also taken into account for practical energy-saving consideration, with α_q and α_p determining their weights in the objective function respectively.

Power flow in voltage regulation adheres to equality constraints derived from the energy conservation of nodal influx and efflux, as:

$$P_{\text{PV}}^i - P_{\text{L}}^i = V^i \sum_{j \in \mathcal{B}^i} \left[V^i G^{ij} - V^j (G^{ij} \cos\theta^{ij} + B^{ij} \sin\theta^{ij}) \right], \quad i \in \mathcal{N} \tag{2}$$

$$Q_{\text{PV}}^i - Q_L^i = V^i \sum_{j \in \mathcal{B}^i} \left[-V^i B^{ij} + V^j (G^{ij} \sin\theta^{ij} + B^{ij} \cos\theta^{ij}) \right], i \in \mathcal{N} \tag{3}$$

where V^i denotes the voltage on bus i, with θ^i being its phase angle, and the difference between phase angles is written as $\theta^{ij} = \theta^i - \theta^j$; G^{ij} and B^{ij} respectively represent the conductance and susceptance on branch lines connecting i and j; while G^{ij} and B^{ij} respectively represent the conductance and susceptance on branch lines connecting i and j. Since this paper mainly focuses on the inverter-based ADN voltage control problem under high PV penetration, active power injection and outflow are considered separately for PVs P_{PV}^i and loads P_{L}^i.

To guarantee the normal operation of ADNs, each bus voltage in the network should be maintained within the range of:

$$0.95\,\text{p.u.} \leqslant V^i \leqslant 1.05\,\text{p.u.}, \quad i \in \mathcal{N} \tag{4}$$

In ADNs, the considered controllable devices comprise DERs and reactive compensation devices, whose roles are to collaborate to satisfy constraints on the bus voltage. In practice, the grid is supposed to meet the demand of individuals, corporations and industries, hence loads cannot participate in the cooperative control to sustain voltage stability. The dynamic conditions considered in this paper for ADNs pertain to the intermittency and uncertainty of PV generation. Considering the small phase shift of the voltages at the two ends of the line, the phase angle difference can be ignored. Multiplying the two ends of (2) and (3) by the branch lines resistance R^{ij} and reactance X^{ij}, respectively, before adding them together leads to:

$$(V^j)^2 - V^i V^j + \left[R^{ij} (P_{\text{L}}^j - P_{\text{PV}}^j) + X^{ij} (Q_{\text{L}}^j - Q_{\text{PV}}^j) \right] = 0 \tag{5}$$

Assuming the voltage magnitude at bus j as the variable and that the equation possesses real roots, a solution must exist, i.e.:

$$V^j = \frac{V^i + \sqrt{(V^i)^2 - 4\left[R^{ij} (P_{\text{L}}^j - P_{\text{PV}}^j) + X^{ij} (Q_{\text{L}}^j - Q_{\text{PV}}^j) \right]}}{2} \tag{6}$$

According to (6), it focuses on P_{PV} and Q_{PV} to maintain voltage safety under dynamic conditions. PVs are chosen as a representative to study the effect of DERs' active curtailment on the nodal voltages within the system, and their constraints can be formulated as:

$$0 \leqslant P_{\text{cur}}^i \leqslant \varepsilon_p \bar{P}_{\text{PV}}^i, \quad i \in \mathcal{G} \tag{7}$$

$$P_{\text{cur}}^i = \bar{P}_{\text{PV}}^i - P_{\text{PV}}^i \tag{8}$$

where $\mathcal{G}$ is the set of distributed generators, namely, the PV set; $\bar{P}_{\text{PV}}^i$ represents the maximum output of PV, and its maximum attenuation factor is $\varepsilon_p \in [0,1]$; and P_{cur}^i equals to the difference between the maximum PV generation and the actual PV output.

PV inverters are capable to generate or absorb reactive power within their capacities, which is one of the optimal choices to optimize the grid voltage profile in a fast-timescale. The capacity constraints can be characterized as:

$$\left|Q_{\text{PV}}^{i}\right| \leqslant \sqrt{(S_{\text{PV}}^{i})^{2}-(P_{\text{PV}}^{i})^{2}},\quad i\in\mathcal{G} \tag{9}$$

In practice, PV generators and inverters are often featured in pairs and therefore share a collective set $\mathcal{G}$. S_{PV}^{i} represents the apparent power of PV, which is typically set to be slightly larger than the maximum capacity of PV [30]. Herein, it considers $S_{\text{PV}}^{i}=1.2\bar{P}_{\max}^{i}$.

Given aforementioned grid states are unavoidably noised in real-world perception and transmission along the coordination chain, perturbations need be taken into account to retain safety and robustness for voltage regulation. The causes of such uncertainty include, but are not limited to, inevitable stochastic measurement errors and outliers, and ubiquitous noise in transmission. Recent studies haven't reached a consensus on the description of these state perturbations [44], [51], [57]–[59]. To simplify the discussion, they are modeled as zero-mean truncated Gaussian distributions $\delta \sim N_{\text{trunc}}(0,\sigma)$, thereby guaranteeing the compactness of the noise set and avoiding unrealistic anomalous perturbations [60]. Here, σ dominates the disturbing magnitude.

III. Proposed Voltage Control Method

In this section, a novel SDRL-based robust voltage control method is proposed that can counteract state perturbations whilst guaranteeing voltage compliance with safety constraints, enhancing the voltage stability of ADNs and reducing energy loss. In the following parts, the voltage regulation task is firstly modeled as a safe Markov game with constraints. To solve the CMDP problem, a novel multi-agent safety reinforcement learning method, i.e., S-MADDPG, is introduced. Afterwards, an anti-perturbation technique of states is incorporated with the SDRL algorithm where both safety and robustness are strengthened with respect to existing voltage regulation approaches. The overall framework of the proposed method is shown in Fig. 1.

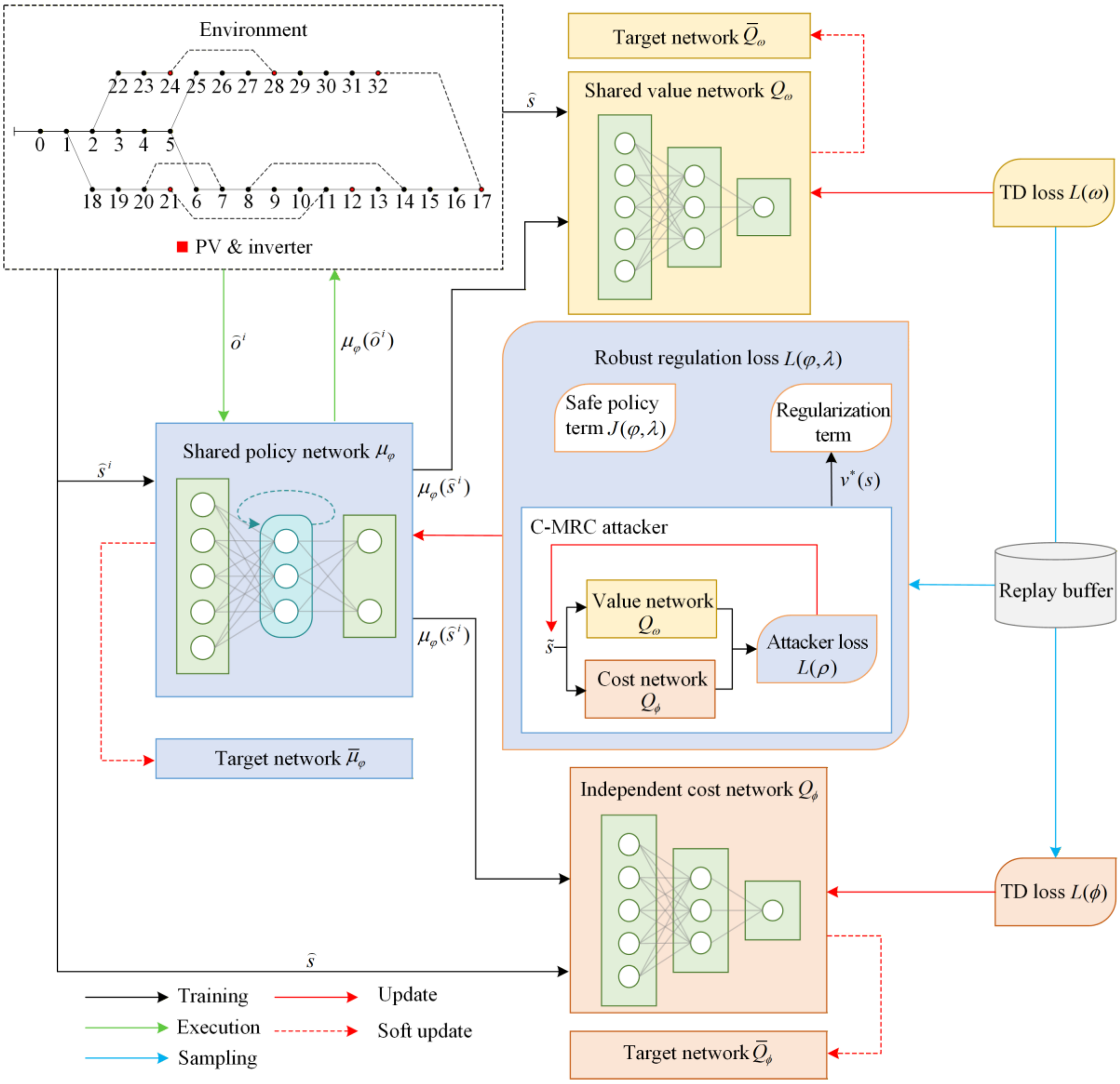

Fig. 1. Proposed robust voltage control framework via SDRL.

A. CMDP Formulation

Voltage control is formulated as a CMDP in multi-agent manner, which is designed to strictly constrain the bus voltage within the safety interval. A standard CMDP can be described as $\langle \mathcal{M}, \mathcal{S}, \mathcal{O}, \mathcal{A}, p, r, c, \bar{c}, \gamma \rangle$. Therein, $\mathcal{M}$ represents the set of agents, which is an entity carrying out environmental interactions and controlling operations; $\mathcal{S}$ denotes the state space of environment, though due to the inevitability of perturbations, actual state values are always unavailable; Hence the observation space $\mathcal{O}$, which stands for the measurements, is adopted in the testing phase in place of $\mathcal{S}$; $\mathcal{A} = \prod_{i \in \mathcal{M}} \mathcal{A}^i$ is the joint action space, encompassing every agent's decisions on each controllable device; The state transition probability $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ shows the probability distribution of the next state s_{t+1} given the state s_t and action a_t in current moment, namely the dynamics of environment; $r, c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote the reward function and the cost function, respectively; $\bar{c}$ is the ceiling of corresponding cost for safety consideration. Typically, the optimization objective of CMDP is to solve the optimal policy π^* over a certain time span T such that the expected return $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t \right]$ is maximized while adhering to the safety constraints of $J_c(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t c_t \right] \leqslant \bar{c}$. Here, $\gamma \in [0,1]$ indicates the time discount factor.

As for the voltage regulation issue, the CMDP components are specifically designated as follows:

Agent: In ADNs, each PV is treated as an independent agent, as the distributed multi-agent control has been implemented in this paper. Thus, the agent set can establish a mapping with the set of DERs indicated as $\mathcal{M} \to \mathcal{G}$.

State & observation space: The state ought to reflect the current profile of global ADN, which is defined as $s_t = (P_{\text{PV}}(t), Q_{\text{PV}}(t), P_L(t), Q_L(t), V(t), \theta(t))$. The global state of real world is revealed using local measurements, also known as observations. Considering the communication expenditure and decision-making cost, observations for each agent are confined to their demarcated regions $\mathcal{R}^i (i \in \mathcal{M})$. These domains have been preordained in alignment with the geographical coordinates of bus nodes, ensuring their non-overlapping spatial delineations. It is noteworthy that each agent belongs to a unique region and each region is allowed to be shared by multiple agents. In practice, such zonal observations are more likely to be contaminated by perturbations. Therefore, the set of local observations is defined as $\left\{ o_t^i = (\tilde{P}_{\text{PV}}^i(t),\ \tilde{Q}_{\text{PV}}^i(t),\ \tilde{P}_L^j(t),\ \tilde{Q}_L^j(t),\ \tilde{V}^j(t),\ \tilde{\theta}^j(t)) \mid i \in \mathcal{M}, j \in \mathcal{R}^i \right\}$ where tildes indicate that observational states are under noising.

Action space: For each agent i, the action space $\mathcal{A}^i$ is specified as the output ratio of the active and reactive power capacities, also denoted as $\left\{ (a_{p,t}^i, a_{q,t}^i) \mid a_{p,t}^i \in \left[1.0 - \varepsilon_p, 1.0\right],\ a_{q,t}^i \in \left[-\varepsilon_q, \varepsilon_q\right] \right\}$. Thus, the actual output of PV generator i and its corresponding inverter can be calculated $P_{\text{PV}}^i(t) = a_{p,t}^i \bar{P}_{\text{PV}}^i(t), Q_{\text{PV}}^i(t) = a_{q,t}^i \sqrt{(S_{\text{PV}}^i)^2 - (P_{\text{PV}}^i(t))^2}$. It is worth mentioning that the whole action space is continuous and its constraints for each dimension are presented in (7) and (9).

Reward function: In Markov games, the design of the reward function is generally in accordance with the optimization objective or its variants. As a result, the cooperative reward is defined as the voltage deviation and reactive power generation of PV inverter, with specifics as:

$$r_t = -\left[\frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \left| V^i(t) - V_{\text{ref}} \right| + \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \left(\alpha_q \left| Q_{\text{PV}}^j(t) \right| + \alpha_p P_{\text{cur}}^j(t) \right) \right] \tag{10}$$

Cost function: In voltage regulation tasks, the cost function is meant to maximally penalize overvoltage violations at arbitrary nodes. Therefore, costs are defined here as a kind of binary function:

$$c_t = \begin{cases} 1, & \psi_{\text{viol},t} \geqslant 1\% \\ 0, & \text{else} \end{cases} \tag{11}$$

where $\psi_{\text{viol},t}$ stands for the ratio of nodes whose voltage violate the safety constraint at time step t, and its upper limit is set to 0.001 by default. Such a ‘hard restriction’ on cost using a binary function has previously been demonstrated to yield promising outcomes for CMDP in other applications [61].

B. S-MADDPG

The proposed algorithm is derived from MADDPG [62], a prevalent framework for multi-agent off-policy reinforcement learning. In contrast to DDPG, its multi-agent variant adopts the CTDE architecture, which implies making decisions leveraging local observations and training policy quality critics according to global states. This distributed decision-making style effectively achieves the globally optimal while minimizing the additional computation and communication burden, which makes it well-suited for addressing the challenge of distributed voltage control tasks in scenarios with a large number of DERs accessing the system. Furthermore, MADDPG has been demonstrated to be one of the state-of-the-art MARL-based methods to address vanilla active voltage control issues [24], which makes it an even more competent baseline.

For each agent, MADDPG provides an actor-critic structure that involves a policy network $\mu^i_\varphi\left(i\in\mathcal{M}\right)$ for decision making and a value network $Q^i_\omega\left(i\in\mathcal{M}\right)$ for assessing these decisions. In practice, μ^i_φ and Q^i_ω are implemented by deep neural networks (DNNs), with φ and ω being the parameters of neurons. Both networks have their corresponding target networks $\bar{\mu}^i_\varphi$ and $\bar{Q}^i_\omega$. Each policy network μ^i_φ executes in a decentralized manner, perceiving only local states s^i and generating the deterministic policy that maximizes the expected return, as:

$$J(\varphi^i)=\mathbb{E}_{s,a\sim\mathcal{D}}\left[Q^i_\mu(s,a)|\ a^i=\mu^i_\varphi(s^i)\right] \tag{12}$$

where $s=(s^1,\cdots,s^M)$ and $a=(a^1,\cdots,a^M)$ denote the respective concatenation of states and actions for all M agents; $\mathcal{D}$ is an experience replay buffer storing agents' dynamic trajectories at each time step and $Q^i_\mu(s,a)=\mathbb{E}_\mu\left[\sum_{t=0}^T\gamma^t r^i(s_t,a_t)|s_0=s,a_0=a\right]$ is an action-value function taking as input the actions and states of all agents, which reflects the expected return of executing action a according to the current strategy μ. Hence, the actor network μ^i_φ ought to update its parameters along the ascending direction of gradient $\nabla_{\varphi^i}J\left(\varphi^i\right)$ during the training phase. There are various ways to approximate Q values [62], and the critic network Q^i_ω is employed in MADDPG. As with DQN, each critic network Q^i_ω updates its parameters using temporal difference (TD) loss:

$$\begin{aligned}L(\omega^i)=\mathbb{E}_{s,\hat{s},a\sim\mathcal{D}}\Big[\Big(r^i(s,a)+\gamma\bar{Q}^i_\omega(\hat{s},\hat{a})-\\ Q^i_\omega(s,a)\Big)^2\mid\hat{a}^i=\bar{\mu}^i_\varphi(\hat{s}^i)\Big]\end{aligned} \tag{13}$$

where $r^i(s,a)+\gamma\bar{Q}^i_\omega(\hat{s},\hat{a})$ represents the expected return after executing action a, which is generally estimated as the objective of Q function. The target networks $\bar{Q}^i_\omega$ and $\bar{\mu}^i_\varphi$ are used to approximate the action-value of next time step, whose parameters are slowly updated with momentum according to the actor-critic networks with the aim of stabilizing the learning process.

When decision-making with constraints is considered in MADDPG, the optimization objective of the policy network changes into:

$$\begin{aligned}&\max_{\varphi^i}J(\varphi^i)=\mathbb{E}_{s,a\sim\mathcal{D}}\left[Q^i_\mu(s,a)\middle|a^i=\mu^i_\varphi(s^i)\right]\\ &\text{s.t.}\quad J_c(\varphi^i)=\mathbb{E}_{s,a\sim\mathcal{D}}\left[Q^i_{c,\mu}(s,a)\middle|a^i=\mu^i_\varphi(s^i)\right]\leqslant\bar{J}^i_c\end{aligned} \tag{14}$$

where $Q^i_{c,\mu}(s,a)=\mathbb{E}_\mu\left[\sum_{t=0}^T\gamma^t c^i(s_t,a_t)|s_0=s,a_0=a\right]$ is the action-cost-value function that measures the discounted expected cost of performing action a; and $\bar{J}^i_c$ is its upper safety limit. Since the optimization problem formulated in (14) is intractable, Lagrange multipliers are introduced to transform it into an unconstrained max-min problem:

$$\max_{\varphi^i}\min_{\lambda^i}J(\varphi^i,\lambda^i)=J(\varphi^i)-\lambda^i\left(J_c(\varphi^i)-\bar{J}^i_c\right) \tag{15}$$

where the Lagrange multiplier $\lambda^i>0$ serves as an adaptable parameter that reveals the current deviation from constraints and provides compensatory adjustments in the reversed direction, thereby ensuring that the voltage remains within the feasible domain throughout the optimization process. In S-MADDPG, λ is updated via gradient descent $\lambda^i\leftarrow\mathrm{ReLU}\left(\lambda^i-\eta_\lambda\nabla_\lambda J_s(\varphi^i)\right)$, where the rectified linear unit (ReLU) activation function is employed to enforce non-negativity. The specific settings for the learning rate η_λ and initial value λ_0 are provided in Section IV.A.

It is noteworthy that the proposed S-MADDPG methodology also incorporates two techniques aimed at enhancing voltage safety and improving computational efficiency. First, resembling to the critic network Q^i_ω, a cost value network Q^i_ϕ and its paired target network $\bar{Q}^i_\phi$ are established to estimate the cost Q function independently, which leads to heightened precision in estimating constraints, and further amplifying the rate of satisfaction. The objective function of the cost Q network is formulated as:

$$\begin{aligned}L(\phi^i)=\mathbb{E}_{s,\hat{s},a\sim\mathcal{D}}\Big[\left(c^i(s,a)+\gamma\bar{Q}^i_\phi(\hat{s},\hat{a})\right)^2-\\ Q^i_\phi(s,a)\mid\hat{a}^i=\bar{\mu}^i_\varphi(\hat{s}^i)\Big]\end{aligned} \tag{16}$$

Second, the utilization of shared network parameters among agents is justified due to the cooperative nature of PVs in voltage control tasks. This technique enhances communication among agents while simultaneously reducing the network scale, thereby increasing the practical relevance of the algorithm. Furthermore, S-MADDPG adopts a deterministic policy, which inherently possesses limited exploration capabilities. Therefore, random noises are introduced into the action sampling during policy learning to augment its exploratory behavior, albeit increasing the risk of voltage violations. To mitigate the risks of voltage violations from aggressive exploration, bounds are enforced on the actions applied to the environment and violations arepenalized through a cost function. This dual approach balances exploration with safety and stability.

C. Robust Regulation Loss

In practical scenarios, the physical state of ADNs is inevitably subject to various perturbations during observation. To address this challenge, a robust regulation loss is proposed, enabling S-MADDPG to attain satisfactory results even in the presence of varying degrees of state disturbances.

Given the inherent uncertainty of perturbation ν, each agent is required to learn and make optimal decisions, accounting for the worst-case scenario. This is crucial to guarantee the robustness of the algorithm. As a result, the objective function of the actor network can be reformulated as:

$$\left(\max_{\varphi}\min_{\lambda}\right)\min_{\nu} J(\varphi,\lambda,\nu)=J(\varphi)-\lambda\left(J_c(\varphi)-\bar{J}_c\right)\Big|_{s=\nu(s)} \tag{17}$$

where $\nu(s)$ indicates the noise-contaminated states. Notwithstanding deriving the optimal deterministic policy μ^* under the optimal perturbation ν^* may present inherent difficulties, it is still possible to establish an upper bound on the expected return loss when facing the worst conditions, as compared to scenarios where perturbations are absent [63]:

$$\begin{aligned}\max_{s\in\mathcal{S}}\left[Q_\mu(s,a)\Big|_{a=\mu(s)}-Q_\mu(\tilde{s},\tilde{a})\Big|_{\tilde{s}=\nu^*(s),\tilde{a}=\mu(\tilde{s})}\right]\leqslant\\ \kappa\max_{s,\tilde{s}\in\mathcal{S}}D\big(\mu(s),\mu(\tilde{s})\big)\end{aligned} \tag{18}$$

where μ represents the regular actor network in the absence of noise; and D measures the deviation between the policy derived from natural states $\mu(s)$ and the contaminated ones $\mu(\tilde{s})$. This theorem suggests that the maximum difference between the Q value under the optimal perturbation and the natural one is bounded by the worst collapse on policies due to the detrimental states. In simpler terms, if one aims to mitigate the impact of state disturbances on the expected return of actions as effectively as possible, it becomes imperative to impose regulations on the maximum deviation during the actor training phase, which ensures the stability of the network's output even in the most adverse conditions. Hence, in order to incorporate robustness into the optimization objective of the policy network, the robust regulation loss is introduced as:

$$L(\varphi,\lambda)=J(\varphi,\lambda)-\beta\sum_{s\in\mathcal{D}}\max_{\nu}\mu(s)-\mu\big(\nu(s)\big)_2 \tag{19}$$

where the l2-normed regularization term is employed to evaluate the variation between policies $D\big(\mu(s),\mu(\tilde{s})\big)$. The advanced S-MADDPG boosted by robust regulation loss is termed RS-MADDPG, which can achieve a potently robust voltage control under the state perturbations in practice. Its training process is elaborated in Algorithm 1.

Algorithm 1 Training Procedures of the Proposed RS-MADDPG

```
Initialize network parameters $\varphi,\omega,\phi$; experience replay buffer $\mathcal{D}$;
Lagrange multiplier $\lambda$; perturbation scale $\rho_s$
Initialize corresponding target network parameters with replications:
$\bar{\varphi}\leftarrow\varphi,\bar{\omega}\leftarrow\omega,\bar{\phi}\leftarrow\phi$
for each episode do
    Reset the environment
    Compute perturbation scale with progressive ascent schedule:
    $\rho_s=\begin{cases}0, if\ \text{cur ep}<\text{start ep}\\ \bar{\rho}_s\left(1-\mathrm{e}^{-5\times\frac{\text{cur ep - start ep}}{\text{duration}}}\right), \text{if cur ep}\geqslant\text{start ep}\end{cases}$
    for each step $t$ do
        Get $s_t^i=\left(P_{\mathrm{PV}}^i(t),Q_{\mathrm{PV}}^i(t),P_L^j(t),Q_L^j(t),V^j(t),\theta^j(t)\right)$, $i\in\mathcal{M}$, $j\in\mathcal{R}^i$
        Encode local states with their agent identities and concatenate as a whole $\hat{s}_t$
        Sample distributed actions $a_t^i=\mu_\varphi(\hat{s}_t)$
        Execute $a_t^i$ in the environment, receiving reward $r_t$, cost $c_t$, and states of next step $s_{t+1}$
        Push transition $\{s_t,a_t,r_t,c_t,s_{t+1}\}$ into buffer $\mathcal{D}$ and pop the oldest one if $\mathcal{D}$ overflowed
        if not updating step then
            continue
        end if
        for each updating step do
            Sample a mini-batch from $\mathcal{D}$
            Obtain optimal perturbed states with C-MRC attacker in Algorithm 2
            Update actor network $\mu_\varphi$ utilizing robust regulation loss in Eq. (19)
            Update value network $Q_\omega$ utilizing TD loss in Eq. (13)
            Update cost value network $Q_\phi$ utilizing TD loss in Eq. (16)
            Update Lagrange multiplier $\lambda$ utilizing robust regulation loss in Eq. (19)
            Update target networks softly: $\bar{\varphi}\leftarrow\tau\varphi+(1-\tau)\bar{\varphi}$,
            $\bar{\omega}\leftarrow\tau\omega+(1-\tau)\bar{\omega},\bar{\phi}\leftarrow\tau\phi+(1-\tau)\bar{\phi}$
        end for
    end for
end for
```

Regarding the problem of solving for the optimal perturbation, a state-of-the-art state adversary known as the combined maximum reward and cost (C-MRC) attacker [64] is employed. This adversary treats the state as a variable and leverages gradient ascent to maximize the reward and cost functions, stealthily enticing the controller to generate policies that violate safety constraints. The C-MRC attacker has demonstrated excellent performance in safe reinforcement learning, making it an ideal choice for generating perturbated states which effectively maximize the regularization term $\big\|\mu(s)-\mu\big(\nu(s)\big)\big\|_2$. The detailed procedure for the C-MRC attacker is specified in Algorithm 2.

Algorithm 2 Operational Procedures of C-MRC Attacker

```
Input: current actor network $\mu_\varphi$, value network $Q_\omega$ and cost value network $Q_\phi$; batched states $s$ and actions $a$; current perturbation scale $\rho_s$; break conditions $\epsilon_L$ and $\epsilon_\rho$
for each episode do
    Output: perturbated states $\tilde{s}$
    Initialize state perturbation variable $\rho$
    for each iteration do
        Compute attacker loss $L(\rho)=\xi_r Q_\omega\big(s,\mu_\varphi(s+\rho)\big)-\xi_c Q_\phi\big(s,\mu_\varphi(s+\rho)\big)$
        Update $\rho:\rho\leftarrow\rho+\eta_\rho\nabla_\rho L(\rho)$
        Clamp $\rho$ within current perturbation scale $[-\rho_s,\rho_s]$
        if $\mathrm{d}L(\rho)<\epsilon_L$ and $\mathrm{d}\rho<\epsilon_\rho$ then
            break
        end if
    end for
    return $\tilde{s}=s+\rho$
```

IV. Numerical Study

This section employs simulation experiments to empirically substantiate the effectiveness of the proposed method in achieving voltage profile stability and resistance against state perturbations. To evaluate its universality, two ADN simulation cases of differing scales are utilized, consisting of mutated IEEE 33-bus [65] and 141-bus [66], respectively, as the environments for the robust SDRL-based voltage control algorithm. By analyzing the voltage control performance of the proposed algorithm under varying degrees of state perturbations and comparing it against benchmark algorithms, we validate its appreciable robustness and superiority.

A. Experimental Setup

The simulation environment in this paper is based on the IEEE 33-bus and 141-bus systems, whose plain versions are derived from MATPOWER. In order to emulate a real-world ADN with high PV penetration while implementing reactive power optimization, 6 and 22 PV generators paired with PV inverters are added to corresponding networks, whose maximum capacities are 0.5 MW and 0.05 MVar, respectively.

The training data, with reference to [24], is selected from real-world Portuguese electricity usage data for loads, and generation data disclosed by Elia group as the active outputs of PVs. Data for each specific node accessing PVs or loads are collected over a 3-year span, sampling distinct real devices at 3-minute intervals. Considering the volatility of physical power equipment, we add Gaussian noise to the PV and load data. In the process of algorithm training and testing, the maximum power output $\overline{P}_{\mathrm{PV}}$ for each solar device is determined using PV data, while ε_p is fixed at 0.7 in order to make use of more active power generation. For safety considerations, ε_q is set to 0.8 and 0.6 in the 33-bus and 141-bus cases, respectively, due to different power profiles between the two scenarios.

For each episode in the reinforcement learning setting, we randomly select a starting timestamp from the 3-year time horizon and take a consecutive series of 240 data to facilitate environmental interactions, simulating the execution of voltage regulation for the ADN within a half day. All proposed algorithms, along with the benchmarks used for comparison, undergo 500 training episodes. The training effectiveness is assessed once in each of the 10 epochs. During the testing phase, each algorithm executes 50 consecutive episodes under varying levels of state perturbation, and the average value is calculated to assess the algorithm's stability and robustness in terms of voltage control. Considering the inherent uncertainty associated with reinforcement learning, each experiment is conducted in parallel five times, utilizing a distinct random seed for each iteration. Some important hyperparameters for the algorithms that emerge later are presented in Table I.

TABLE I
HYPERPARAMETERS OF UPCOMING ALGORITHMS

Algorithm	Parameter	value
Collective	Policy network	RNN
	(Cost) Value network	MLP
	Hidden layer size	64
	Optimizer	RMSprop
	η_φ,η_ω	0.0001
	τ	0.1
	Total episodes	500
	Total steps per episode	240
	Replay buffer size	5000
	α_p,α_q	0.1
	γ	0.99
MAPPDG	Network update frequency (step)	60
	Target update frequency (step)	120
S-MAPPDG	λ_0	0.4
	η_λ	0.0001
	$\overline{J}_c$	0
RS-MAPPDG	$\overline{\rho}_s$	0.04
	Start episode	100
	Duration (episode)	100
	ϵ_L,ϵ_ρ	0.0001
	ρ_0	0
	η_ρ	0.01
	ξ_r,ξ_c	−0.5
	β	1

It is worth mentioning that all algorithms in this section are implemented under the Pandapower and Pytorch frameworks in Python. This work is done on a workstation with 128 GB memory, Intel Xeon(R) Gold 5220 R CPU, and Quadro RTX6000 GPU.

B. Convergence and Effectiveness of the Proposed Algorithm

Figures 2 and 3 depict the mean reward and cost performance of three algorithms, namely, vanilla MADDPG, S-MADDPG, and the proposed RS-MADDPG, during the training phase for two distinct case scales, with the shading as the 95% confidence interval around the mean. Both the indicated reward and cost values represent the averages over 240 consecutive environmental feedbacks within each episode. From Fig. 2 and Fig. 3, it is evident that in all scenarios, the rewards of the three algorithms increase sharply, while the associated costs decline significantly as the training progresses. All algorithms exhibit rapid convergence towards their respective optimal values at approximately 100 episodes. These results demonstrate that the newly introduced components enables the MADDPG algorithm to maintain strong convergence performance. As observed from the training cost curves, both control algorithms modeled by CMDP exhibit substantially lower values compared to the conventional MADDPG algorithm. This substantiates that incorporating the Lagrange multiplier further enhances the algorithm's capacity to stabilize nodal voltages of ADN and facilitate improved voltage regulation, even when the voltage deviation is confined within the reward function. Nevertheless, due to the absence of perturbations during training, the disparity in costs between S-MADDPG and RS-MADDPG is not significant, as demonstrated in subsequent perturbation-free test experiments. Regarding the performance of training rewards, there are no notable differences among those three algorithms. This signifies that the proposed innovations in this paper can further augment the voltage steadiness and robustness without compromising the network's optimization prowess, thus presenting excellent prospects for practical applications.

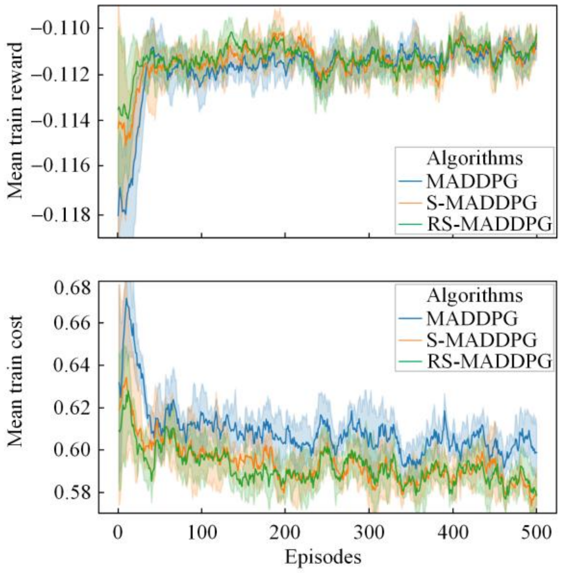

Fig. 2. Training evolution of the proposed RS-MADDPG and its baselines (MADDPG, S-MADDPG) in the 33-bus case.

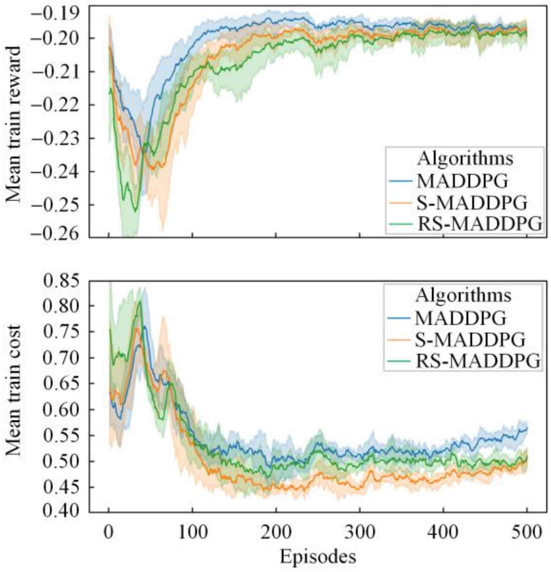

Fig. 3. Training evolution of the proposed RS-MADDPG and its baselines (MADDPG, S-MADDPG) in the 141-bus case.

To validate the efficacy of the proposed algorithm in achieving practical voltage control, a series of tests on well-trained models across 2 scenarios are conducted, encompassing both unperturbed and perturbed conditions. A conventional model-based control method is also selected to provide a more convincing comparison. This optimization-based algorithm follows the objective function and constraints mentioned in Section II to build a voltage control model of ADN, which is then relaxed to a standard second-order cone programming (SOCP) problem [67] and finally solved with the GUROBI solver. The corresponding outcomes are presented in Table II and Table III, respectively, demonstrating the empirical evidence for the algorithm's effectiveness and robustness. The performance of voltage control algorithms is assessed by two essential metrics: power loss (PL) and controllable rate (CR) [24]. The PL quantifies the mean energy loss over entire buses, reflecting the algorithm's validity on minimizing the primary objective of optimization. On the other hand, CR evaluates the safety by calculating the average ratio of time steps where all buses operate within the prescribed voltage limits in each episode. In this paper, the reported values for both metrics are step-level averages over the entire range of episodes, and except for the model-based approach, the means and standard deviations for five parallel tests with distinct seeds are summarized in the tables, providing a reliable assessment of the consistency.

TABLE II
TEST PERFORMANCE WITHOUT STATE PERTURBATIONS

Test case	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
33-bus	Model-based	**0.0475**		0.9941	
	MADDPG	0.0546	0.0046	0.9437	0.0206
	S-MADDPG	0.0594	0.0070	0.9917	0.0021
	RS-MADDPG	0.0620	0.0084	0.9861	0.0056
141-bus	Model-based	**0.4605**		0.9998	
	MADDPG	0.6417	0.0933	0.9987	0.0012
	S-MADDPG	0.8196	0.0730	**1.0000**	0.0000
	RS-MADDPG	0.7693	0.0949	0.9979	0.0007

TABLE III
TEST PERFORMANCE WITH STATE PERTURBATIONS

Test case	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
33-bus	Model-based	0.0585		0.8812	
	MADDPG	0.0560	0.0016	0.8749	0.0237
	S-MADDPG	0.0621	0.0035	0.8897	0.0309
	RS-MADDPG	**0.0510**	0.0040	**0.9345**	0.0224
141-bus	Model-based	0.8302		0.9244	
	MADDPG	**0.5887**	0.0259	0.8462	0.1301
	S-MADDPG	0.7492	0.1164	0.9380	0.0522
	RS-MADDPG	0.6239	0.0643	**0.9906**	0.0053

As noticed in Table II, the model-based approach reigns over all comprehensive test results, whether in terms of PL or CR, in the complete spectrum of scenarios. This demonstrates that traditional optimization algorithms continue to remain a preferred choice for addressing the voltage control problem, without considering model uncertainty and implementation efficiency. Nevertheless, in real-world contexts, the significance of two aforementioned issues cannot be underestimated. Thus, the employment of DRL has gained ascendancy within contemporary researches on voltage control. As for model-free methods, the voltage control capability and stability of S-MADDPG are predominant in various scenarios, aligned with the model-based regulation, which serves as a compelling evidence for the strength of the proposed SDRL framework in reducing the voltage violation rate within the network and consequently bolsters the operational safeness of ADNs. On the other hand, considering the conflicting nature of high voltage security and low line loss according to the power flow equations of the distribution network, the less stringent voltage constraints of the vanilla MADDPG lead to superior line loss regulation. However, the robust safety method doesn't demonstrate a satisfactory performance on the integrated voltage control ability, despite its PL and CR do not differ much from those two above. This phenomenon is attributable to the fact that robust regulation loss prioritizes the attainment of solutions optimized for worst-case scenarios, which is challenging to outperform under ideal test conditions that are completely free of perturbations.

In order to ascertain the efficacy of the proposed voltage control method in effectively countering state perturbations, truncated Gaussian noises are introduced into observations as representations of the state susceptible to noise disturbances during testing. Taking into account the unit diversity of different measured variables, the perturbations are realized via multiplying state values by the noise factor, denoted as $o = \tilde{s} = (1+\delta)\cdot s, \delta \sim N_{\text{trunc}}(0,\sigma)$. In the validity verification, $\sigma = 3.0$ is exclusively established and the ensuing test outcomes are documented in Table III. Evidently, except for a slight disadvantage on PL in the 141-bus grid, RS-MADDPG demonstrates exceptional proficiency in stabilizing nodal voltage and optimizing power loss, surpassing other benchmark algorithms in all test cases. Upon comparing the data presented in Table II, it is seen that, in the presence of a significant state perturbation, RS-MADDPG demonstrates a level of performance nearly equivalent to that of MADDPG in its unaltered, natural state, for both two key metrics related to voltage control. This outcome effectively proves that the proposed robust regulation loss genuinely empowers the SDRL-based voltage controller to acquire more sound strategies, thereby enabling it to achieve robust regulation performance in the face of state perturbations.

C. Robustness Analysis

To enhance the comprehensiveness of the robustness test and demonstrate the control performance of both the proposed algorithm and its benchmarks under varying degrees of state perturbations, 5 stepwise levels of perturbation patterns are selected for optimization tests i.e., $\sigma \sim \{0.5, 1.0, 1.5, 2.0, 2.5\}$. It is worth noting that, apart from the distinct levels of perturbation, all other parameters remained constant throughout the experiments. These tests are conducted on both the 33-bus and 141-bus networks, whose outcomes are outlined in Tables IV and V.

TABLE IV
ROBUST ANALYSIS UNDER VARYING STATE PERTURBATIONS (33-BUS)

Test case		33-bus			
Noise scale	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
$\sigma = 0.5$	Model-based	**0.0484**		0.9750	
	MADDPG	0.0549	0.0039	0.9381	0.0189
	S-MADDPG	0.0599	0.0059	**0.9868**	0.0024
	RS-MADDPG	0.0596	0.0083	0.9818	0.0070
$\sigma = 1.0$	Model-based	**0.0428**		0.9487	
	MADDPG	0.0549	0.0022	0.9206	0.0183
	S-MADDPG	0.0609	0.0041	0.9611	0.0104
	RS-MADDPG	0.0554	0.0067	**0.9678**	0.0097
$\sigma = 1.5$	Model-based	**0.0471**		0.9285	
	MADDPG	0.0551	0.0010	0.9015	0.0202
	S-MADDPG	0.0615	0.0032	0.9316	0.0196
	RS-MADDPG	0.0529	0.0054	**0.9541**	0.0133
$\sigma = 2.0$	Model-based	0.0529		0.9055	
	MADDPG	0.0554	0.0008	0.8898	0.0219
	S-MADDPG	0.0618	0.0031	0.9127	0.0257
	RS-MADDPG	**0.0518**	0.0047	**0.9451**	0.0168
$\sigma = 2.5$	Model-based	0.0593		0.8832	
	MADDPG	0.0557	0.0012	0.8812	0.0232
	S-MADDPG	0.0619	0.0033	0.8993	0.0284
	RS-MADDPG	**0.0513**	0.0043	**0.9390**	0.0199

TABLE V
ROBUST ANALYSIS UNDER VARYING STATE PERTURBATIONS (141-BUS)

Test case		141-bus			
Noise scale	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
$\sigma = 0.5$	Model-based	**0.4673**		0.9973	
	MADDPG	0.6302	0.0867	0.9941	0.0077
	S-MADDPG	0.8096	0.0766	**0.9998**	0.0001
	RS-MADDPG	0.7526	0.0869	0.9977	0.0008
$\sigma = 1.0$	Model-based	**0.4825**		0.9829	
	MADDPG	0.6081	0.0696	0.9586	0.0496
	S-MADDPG	0.7889	0.0973	0.9962	0.0028
	RS-MADDPG	0.7078	0.0729	**0.9968**	0.0014
$\sigma = 1.5$	Model-based	**0.5404**		0.9717	
	MADDPG	0.5966	0.0527	0.9133	0.0888
	S-MADDPG	0.7713	0.1129	0.9787	0.0176
	RS-MADDPG	0.6712	0.0666	**0.9953**	0.0020
$\sigma = 2.0$	Model-based	0.6083		0.9579	
	MADDPG	**0.5921**	0.0410	0.8827	0.1096
	S-MADDPG	0.7612	0.1174	0.9611	0.0333
	RS-MADDPG	0.6495	0.0648	**0.9937**	0.0031
$\sigma = 2.5$	Model-based	0.7021		0.9466	
	MADDPG	**0.5899**	0.0323	0.8622	0.1220
	S-MADDPG	0.7543	0.1177	0.9479	0.0441
	RS-MADDPG	0.6347	0.0643	**0.9920**	0.0044

As shown, under the condition of slight disturbance ($\sigma = 0.5$), S-MADDPG demonstrates optimal voltage control performance, which underscores the inherent anti-disturbance capabilities embedded within the proposed CMDP architecture. As the level of noise intensity grows further, the proposed RS-MADDPG progressively exhibits heightened robustness, effectively mitigating deviations from voltage constraints. Even in the severe circumstances, i.e., $\sigma = 2.5$, the robust regulation algorithm still manages to maintain the CR with no less than 93% and 99% in the 33-bus and 141-bus test cases, respectively. In terms of power loss performance, the model-based algorithm has an intrinsic advantage, particularly when observations relatively approximate the true system states. Model-free approaches, particularly RS-MADDPG, regain their superiority only when σ exceeds a value greater than 2. Although the proposed algorithm may not offer optimality in line-loss optimization in the 141-bus case, the marginal discrepancy observed in PL may seem inconsequential when considering the substantial comparative advantage in CR. Hence, the proposed RS-MADDPG shows adequate robustness performance on voltage control under varying state disturbances. This robustness enhances the operational safety and adaptability for real-world ADNs tackling the fickle realities.

D. Performance Comparisons with Other Methods

Given the widespread utilization of diverse MARL-based optimal control techniques in power systems, the significance of the generalizability of our proposed robust safety framework cannot be neglected. Meanwhile, suppose the assimilation of the proposed innovations into these algorithms, whether the proposed approach, RS-MADDPG, is the most appropriate for addressing the ADN voltage control problem under state perturbations requires further investigation. Hence, in order to comprehensively evaluate algorithm performance, we selected two alternative approaches, namely MATD3 [68] and MAPPO [69], for further examination. Notably, the MATD3 represents a more contemporary off-policy strategy in comparison to MADDPG, while the MAPPO stands as one of the most prevailing on-policy algorithms presently employed. Both algorithms coalesce their own policy losses into (19) to utilize the proposed CMDP architecture and robust regulation loss, with other components immutable. All experiments are conducted under the two distinct scenarios, characterized by the 33 buses and 141 buses, while striving to maintain consistency in training hyperparameters and settings as closely as possible to those employed for RS-MADDPG. To maintain coherence, the new algorithms are named RS-MATD3 and RS-MAPPO respectively.

As depicted in Figs. 4 and 5, both MADDPG and MATD3 trained within the framework proposed in this paper have exhibited promising outcomes. However, RS-MAPPO exhibits a notable inferiority in terms of reward optimization compared to the other two benchmarks, which is contradicted with the findings in general game environments [69]. Additionally, it manifests the slowest convergence speed, accompanied by the most fluctuating training performance, both on reward and cost metrics. These tend to be attributed to its conservative policy updating scheme, which proves challenging to adapt to the intricate dynamics and uncertainties inherent in the ADN environment [24].

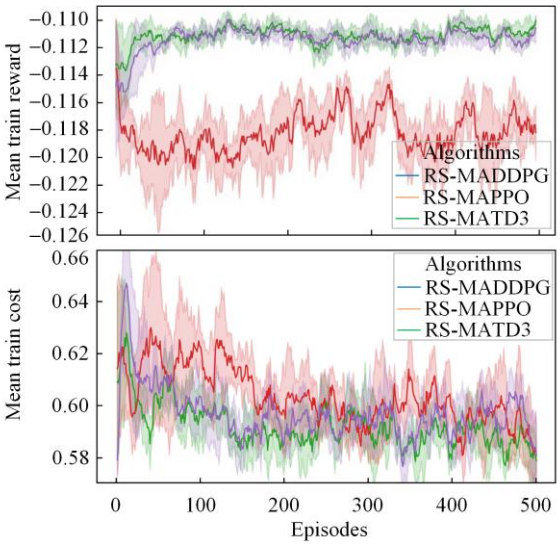

Fig. 4. Training evolution of the proposed RS-MADDPG and other benchmarks (RS-MAPPO, RS-MATD3) in 33-bus case.

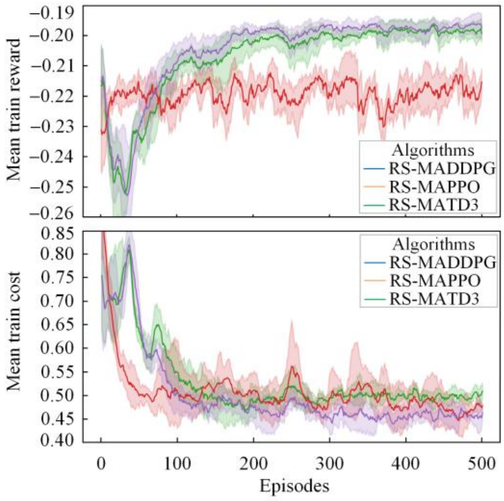

Fig. 5. Training evolution of the proposed RS-MADDPG and other benchmarks (RS-MAPPO, RS-MATD3) in 141-bus case.

Furthermore, considering that there also exist other solutions to CMDP, comparative experiments are conducted with the classical Lyapunov-based SDRL. Lyapunov function constructed on an initial safe strategy ensures that subsequent policy updates remain within a safety set that encompasses the optimal policy, thereby guaranteeing that the cumulative cost incurred under such policies does not exceed a predefined threshold [70]. The robustness test results of the aforementioned four algorithms, along with their respective baselines, are outlined in Tables VI and VII. According to the comprehensive robustness analyses of the RS-MADDPG method previously, three typical levels of state perturbation i.e., $\sigma \sim \{1.0, 2.0, 3.0\}$, are identified, representing mild, moderate and severe levels of perturbation, respectively. Apart from this, all other experimental settings have remained constant throughout the entire test.

TABLE VI
TEST PERFORMANCE COMPARISON WITH OTHER MARL-BASED METHODS (33-BUS)

Test case		33-bus			
Noise scale	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
$\sigma = 1.0$	MADDPG	0.0549	0.0022	0.9206	0.0183
	RS-MADDPG	0.0554	0.0067	**0.9678**	0.0097
	MAPPO	0.1463	0.0496	0.7901	0.1342
	RS-MAPPO	0.1123	0.0301	0.8978	0.0712
	MATD3	0.0499	0.0031	0.8921	0.0467
	RS-MATD3	0.0526	0.0029	0.9518	0.0214
	Lyapuov-based	0.0658	0.0095	0.9585	0.0231
$\sigma = 2.0$	MADDPG	0.0554	0.0008	0.8898	0.0219
	RS-MADDPG	0.0518	0.0047	0.9451	0.0168
	MAPPO	0.1300	0.0441	0.7862	0.1230
	RS-MAPPO	0.1014	0.0279	0.8672	0.0917
	MATD3	0.0544	0.0080	0.8429	0.0975
	RS-MATD3	0.0519	0.0031	0.9345	0.0227
	Lyapuov-based	0.0875	0.0112	0.9254	0.0522
$\sigma = 3.0$	MADDPG	0.0560	0.0016	0.8749	0.0237
	RS-MADDPG	**0.0510**	0.0040	**0.9345**	0.0224
	MAPPO	0.1248	0.0398	0.7805	0.1221
	RS-MAPPO	0.0982	0.0260	0.8506	0.1094
	MATD3	0.0569	0.0113	0.8210	0.1185
	RS-MATD3	0.0520	0.0034	0.9260	0.0240
	Lyapuov-based	0.1396	0.0178	0.8869	0.1273

TABLE VII
TEST PERFORMANCE COMPARISON WITH OTHER MARL-BASED METHODS (141-BUS)

Test case		141-bus			
Noise scale	Algorithm	PL (MW)		CR	
		Mean	Std.	Mean	Std.
$\sigma = 1.0$	MADDPG	**0.6081**	0.0696	0.9586	0.0496
	RS-MADDPG	0.7078	0.0729	**0.9968**	0.0014
	MAPPO	1.1178	0.5129	0.8653	0.0626
	RS-MAPPO	1.2388	0.2905	0.9842	0.0142
	MATD3	0.9488	0.2267	0.9487	0.0666
	RS-MATD3	1.1468	0.3558	0.9923	0.0117
	Lyapuov-based	0.9858	0.2978	0.9577	0.0334
$\sigma = 2.0$	MADDPG	**0.5921**	0.0410	0.8827	0.1096
	RS-MADDPG	0.6495	0.0648	**0.9937**	0.0031
	MAPPO	1.0864	0.4672	0.8485	0.0742
	RS-MAPPO	1.2631	0.3483	0.9681	0.0145
	MATD3	0.8936	0.1647	0.9110	0.0936
	RS-MATD3	1.1115	0.4084	0.9795	0.0325
	Lyapuov-based	1.0645	0.3977	0.9164	0.0683
$\sigma = 3.0$	MADDPG	**0.5887**	0.0259	0.8462	0.1301
	RS-MADDPG	0.6239	0.0643	**0.9906**	0.0053
	MAPPO	1.0670	0.4486	0.8411	0.0804
	RS-MAPPO	1.2693	0.3758	0.9581	0.0231
	MATD3	0.8748	0.1381	0.8968	0.0995
	RS-MATD3	1.0848	0.4232	0.9684	0.0406
	Lyapuov-based	1.2089	0.3893	0.8805	0.0547

First, when comparing the impact of the proposed robust safety framework on each algorithm under state perturbations, it can be observed that the performances of voltage stability among all algorithms are significantly improved across entire scenarios. Moreover, the regulation of power loss in those algorithms exhibits marginal enhancement in the 33-bus case, and does not deviate significantly from their vanilla benchmarks in the 141-bus setting. These outcomes indicate that the CMDP formulation and robust regulation loss introduced in this paper possess decent generalization capability which effectively adapt to a wide range of MARL-based algorithms, thereby promoting their safety and robustness on voltage control tasks of ADN.

Tables VI and VII highlight the optimal test performance of each metrics within the same scenario. It is evident that in both the 33-bus and 144-bus cases under varying perturbations, RS-MADDPG demonstrates the most elite performance with regard to the voltage controllable ratio. Upon examining the power loss performance, we notice that MATD3 emerges as the front-runner in line loss performance under weak perturbations, which is ascribed to its dual-critic network design that effectively mitigates overestimation bias within the 33-bus network [68]. However, as the noise gradually intensifies, RS-MADDPG progressively demonstrates robust competitiveness, and even slightly outperforms RS-MATD3. Additionally, despite the remarkable performance demonstrated by Lyapunov functions in the realm of SDRL, it exhibits certain disadvantages under progressively intensifying state perturbations. This observation underscores the suitability of the proposed robust safety architecture and MADDPG. In the 141-bus scenario, despite RS-MADDPG exhibiting a slightly higher PL compared to its vanilla benchmark, its superior voltage stability is far better than other algorithms across varying degrees of perturbation. In summary, RS-MADDPG demonstrates optimal voltage regulation in the majority of cases and excels in achieving integrated optimization in complex scenarios characterized by high PV penetrations.

V. Conclusion

This paper introduces a robust voltage control method for ADNs that guarantees the safety of nodal voltages and mitigates line loss under state perturbations in a real-time, model-free manner. The proposed method formulates the voltage control problem as a safety Markov game, leverages MADDPG as a basis, and incorporates voltage constraints as independent costs involved in interactions between agents and the environment during training. Moreover, a robust regulation loss is engaged to enhance the robustness on control performance in the presence of state perturbations. Extensive tests conducted on the IEEE 33-bus and 141-bus cases demonstrate the exceptional combined performance on controllable rate of nodal voltages and gross power loss achieved by the proposed RS-MADDPG compared to its benchmarks, across various degrees of state perturbation. The results affirm the robust efficacy and superiority of the proposed approach in practical voltage control optimization. Furthermore, the robust safety framework introduced in this paper also exhibits potential for generalization to other MARL-based algorithms. In future work, the extended application of RS-MADDPG in three-phase unbalanced distribution networks will be further investigated. In addition to state perturbations, the ‘worst-case’ studies of SDRL-based voltage control algorithms in other aspects of and, such as topology, network parameters, uncertainties of PV generation and loads, are the next intriguing point. The aim is to build a robust synthesis that can resolve most of the uncertainty problems on realistic distribution network optimization.

Acknowledgment

Not applicable.

Authors’ Contributions

Meng Tian: writing original draft, methodology, funding acquisition, and conceptualization. Xiaoxu Li: writing review & editing, software, and investigation. Ziyang Zhu: writing review & editing, validation, and funding acquisition. Zhengcheng Dong: writing original draft, validation, software, and investigation. Li Gong: visualization, validation, and software. Jingang Lai: methodology and conceptualization. All authors read and approved the final manuscript.

Funding

This work is supported in part by the National Natural Science Foundation of China (No. 52177109) and Key R&D Program of Hubei Province, China (No. 2020BAB109).

Availability of Data and Materials

Not applicable.

Declarations

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Authors’ Information

Meng Tian received the B.S. degree and the Ph.D. degree from the School of Electronic Information School, Wuhan University, Wuhan, China, in 2011 and 2016, respectively. From 2016 to 2022, he was a postdoctoral research associate/ an associate research fellow with Wuhan University. He was a Visiting Scholar with Southern Methodist University in 2017 and 2018. He is currently an Associate Professor with the School of Automation, Wuhan University of Technology. His research interests include resilience of cyber-physical power systems and the application of deep reinforcement learning in power systems.

Xiaoxu Li received the B.S. degree from the College of Optoelectronic Engineering, Chongqing University, Chongqing, China, in 2022, and the M.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2025. His research interest includes the application of deep reinforcement learning in power systems.

Ziyang Zhu received the B.S. degree and the M.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2021 and 2024, respectively. His research interest includes the application of deep learning techniques in power system and its automation.

Zhengcheng Dong received the B.S. degree from the School of Electric Power, North China University of Water Resources and Electric Power, Zhengzhou, China, in 2011, and the M.S. and Ph.D. degrees from the Electronic Information School and the School of Power and Machinery, Wuhan University, Wuhan, China, in 2013 and 2016, respectively. From 2016 to 2019, he was a postdoctoral research associate with Wuhan University. He is currently an associate professor with the School of Automation, Wuhan University of Technology. His research interests include modeling of cyber-physical systems, planning and optimization of low-carbon integrated energy systems and vulnerability analysis of interdependent networks.

Li Gong received the B.S. degree from the College of Instrumentation & Electrical Engineering, Jilin University, Changchun, China, in 2017, and the Ph.D. degree from the Electronic Information School, Wuhan University, Wuhan, China in 2025. His research interest includes intelligent diagnosis of electrical equipment, decomposition of power marketing operations, and restoration of urban distribution networks.

Jingang Lai received the Ph.D. degree in control science and engineering from Wuhan University, Wuhan, China, in 2016. In 2015, he was a Joint Ph.D. student with the School of Electrical and Computer Engineering, RMIT University, Melbourne, VIC, Australia, where he was a research fellow and an Honorary Principal Research Fellow with the School of Engineering. From 2019 to 2021, he was a Humboldt research fellow and a guest professor with the E.ON Energy Research Center, RWTH Aachen University, Aachen, Germany. He is currently a professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include brain-inspired intelligence and swarm intelligence for microgirds, distributed renewable energy systems, and cyber-physical-social systems.

REFERENCES

[1] Q. Hou, N. Zhang, and E. Du *et al.*, "Probabilistic duck curve in high PV penetration power system: concept, modeling, and empirical analysis in China," *Applied Energy*, vol. 242, pp. 205-215, Mar. 2019.

[2] S. Rahman, S. Saha, and M. Haque *et al.*, "A framework to assess voltage stability of power grids with high penetration of solar PV systems," *International Journal of Electrical Power & Energy Systems*, vol. 139, Dec. 2022.

[3] L. Liu, Y. Zhao, and D. Chang *et al.*, "Prediction of short-term PV power output and uncertainty analysis," *Applied Energy*, vol. 228, pp. 700-711, Dec. 2018.

[4] M. D. Hraiz, J. A. M. García, and R. J. Castañeda *et al.*, "Optimal PV size and location to reduce active power losses while achieving very high penetration level with improvement in voltage profile using modified jaya algorithm," *IEEE Journal of Photovoltaics*, vol. 10, no. 4, pp. 1166-1174, Jul. 2020.

[5] M. Zeraati, M. E. H. Golshan, and J. M. Guerrero, "A consensus-based cooperative control of PEV battery and PV active power curtailment for voltage regulation in distribution networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 670-680, Jan. 2017.

[6] Y. Chai, L. Guo, and C. Wang *et al.*, "Network partition and voltage coordination control for distribution networks with high penetration of distributed PV units," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3396-3407, May 2018.

[7] M. R. Jafari, M. Parniani, and M. H. Ravanji, "Decentralized control of OLTC and PV inverters for voltage regulation in radial distribution networks with high PV penetration," *IEEE Transactions on Power Delivery*, vol. 37, no. 6, pp. 4827-4837, Dec. 2022.

[8] S. Wang, Y. Dong, and L. Wu *et al.*, "Interval overvoltage risk based PV hosting capacity evaluation considering PV and load uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2709-2721, May 2019.

[9] P. Srivastava, R. Haider, and V. J. Nair *et al.*, "Voltage regulation in distribution grids: a survey," *Annual Reviews in Control*, vol. 55, pp. 165-181, Jan. 2023.

[10] Q. Zhang, Y. Zeng, and Y. Hu *et al.*, "Droop-free distributed cooperative control framework for multisource parallel in seaport DC micro-grid," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4231-4244, Nov. 2022.

[11] S. M. Mohiuddin and J. Qi, "Droop-free distributed control for ac microgrids with precisely regulated voltage variance and admissible voltage profile guarantees," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 1956-1967, May 2019.

[12] Y. Liu, Z. Li, and J. Zhao, "Safety-constrained stagewise optimization of droop control parameters for isolated microgrids," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 77-88, Jan. 2024.

[13] X. Sun, J. Qiu, and Y. Tao *et al.*, "A multi-mode data-driven volt/var control strategy with conservation voltage reduction in active distribution networks," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 1073-1085, Apr. 2022.

[14] M. Tahir, M. E. Nassar, and R. El-Shatshat *et al.*, "A review of volt/var control techniques in passive and active power distribution networks," in *2016 IEEE Smart Energy Grid Engineering* (*SEGE*), Oshawa, Canada, Aug. 2016, pp. 57-63.

[15] A. Majumdar, Y. P. Agalgaonkar, and B. C. Pal *et al.*, "Centralized volt-var optimization strat- egy considering malicious attack on distributed energy resources control," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 1, pp. 148-156, Jan. 2017.

[16] Y. Ju, Z. Zhang, and W. Wu *et al.*, "A bi-level consensus ADMM-based fully distributed inverter-based volt/var control method for active distribution networks," *IEEE Transactions on Power Systems*, vol. 37, no. 1, pp. 476-487, Jan. 2021.

[17] D. Gebbran, S. Mhanna, and Y. Ma *et al.*, "Fair coordination of distributed energy resources with volt-var control and PV curtailment," *Applied Energy*, vol. 286, Oct. 2021.

[18] Z. Wang, Y. Wang, and G. Liu *et al.*, "Fast distributed voltage control for PV generation clusters based on approximate newton method," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 612-622, Jan. 2020.

[19] C. Hu, X. Zhang, and Q. Wu, "Collaborative active and reactive power control of DERS for voltage regulation and frequency support by distributed event-triggered heavy ball method," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3804-3815, Sept. 2023.

[20] H. Liu, W. Wu, and Y. Wang, "Bi-level off-policy reinforcement learning for two-timescale volt/var control in active distribution networks," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 385-395, Jan. 2022.

[21] D. Hu, Z. Ye, and Y. Gao *et al.*, "Multi-agent deep reinforcement learning for voltage control with co-

ordinated active and reactive power optimization," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4873-4886, Nov. 2022.

[22] Y. Yuan, K. Dehghanpour, and Z. Wang *et al.*, "A joint distribution system state estimation framework via deep actor-critic learning method," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 796-806, Jan. 2022.

[23] J. Xie and W. Sun, "Distributional deep reinforcement learning-based emergency frequency control," *IEEE Transactions on Power Systems*, vol. 37, no. 4, pp. 2720-2730, Jul. 2021.

[24] J. Wang, W. Xu, and Y. Gu *et al.*, "Multi-agent reinforcement learning for active voltage control on power distribution networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3271-3284, 2021.

[25] X. Chen, G. Qu, and Y. Tang *et al.*, "Reinforce- ment learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2935-2958, Jul. 2022.

[26] K. Xiong, D. Cao, and G. Zhang *et al.*, "Coordinated volt/var control for photovoltaic inverters: a soft actor-critic enhanced droop control approach," *International Journal of Electrical Power & Energy Systems*, vol. 149, Nov. 2023.

[27] Y. Huo, P. Li, and H. Ji *et al.*, "Data-driven coordinated voltage control method of distribution networks with high DG penetration," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1543-1557, Mar. 2022.

[28] M. Wang, M. Feng, and W. Zhou *et al.*, "Stabilizing voltage in power distribution networks via multi-agent reinforcement learning with transformer," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1899-1909.

[29] D. Cao, J. Zhao, and W. Hu *et al.*, "Attention enabled multi-agent DRL for decentralized volt-var control of active distribution system using PV inverters and SVCs," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1582-1592, Jun. 2021.

[30] D. Cao, J. Zhao, and W. Hu *et al.*, "Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning," *Applied Energy*, vol. 306, 2022.

[31] Y. Zhang, X. Wang, and J. Wang *et al.*, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 361-371, Jan. 2020.

[32] S. Wang, J. Duan, and D. Shi *et al.*, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4644-4654, Nov. 2020.

[33] D. Cao, W. Hu, and J. Zhao *et al.*, "A multi-agent deep reinforcement learning based voltage regulation using coordinated pv inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[34] Y. Wang, D. Qiu, and G. Strbac *et al.*, "Coordinated electric vehicle active and reactive power control for active distribution networks," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1611-1622, Feb. 2022.

[35] X. Zhang, Y. Liu, and J. Duan *et al.*, "Ddpg-based multi-agent framework for SVC tuning in urban power grid with renewable energy resources," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5465-5475, Nov. 2021.

[36] D. Cao, J. Zhao, and W. Hu *et al.*, "Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4137-4150, Nov. 2021.

[37] D. Cao, J. Zhao, and W. Hu *et al.*, "Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems," *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 149-165, Jan. 2021.

[38] Q. Zhang, K. Dehghanpour, and Z. Wang *et al.*, "Multi-agent safe policy learning for power management of networked microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1048-1062, Mar. 2020.

[39] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 1860-1872, May 2022.

[40] 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.

[41] W. Wang, N. Yu, and 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, 2019.

[42] H. T. Nguyen and D.-H. Choi, "Three-stage inverter-based peak shaving and volt-var control in active distribution networks using online safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 3266-3277, Jul. 2022.

[43] J. Zhao and L. Mili, "A framework for robust hybrid state estimation with unknown measurement noise statistics," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1866-1875, May 2017.

[44] J. Liu and Z. Li, "Robust expectation-maximization-based secondary voltage control scheme considering stochastic measurement error," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2958-2961, May 2023.

[45] M. A. Putratama, R. Rigo-Mariani, and V. Debusschere *et al.*, "Mitigation of grid parameter uncertainties for the steady-state operation of a model-based voltage controller in distribution systems," *Electric Power Systems Research*, vol. 218, 2023.

[46] A. Petrusev, M. A. Putratama, and R. Rigo-Mariani *et al.*, "Reinforcement learning for robust voltage control in distribution grids under uncertainties," *Sustainable Energy, Grids and Networks*, vol. 33, 2023.

[47] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2037-2047, May 2020.

[48] Y. Xie, L. Liu, and Q. Wu *et al.*, "Robust model predictive control based voltage regulation method for a distribution system with renewable energy sources and energy storage systems," *International Journal of Electrical Power & Energy Systems*, vol. 118, 2020.

[49] H. Liu, C. Zhang, and Q. Chai *et al.*, "Robust regional coordination of inverter-based volt/var control via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5420-5433, Nov. 2021.

[50] N. Daratha, B. Das, and J. Sharma, "Robust voltage regulation in unbalanced radial distribution system under uncertainty of distributed generation and loads," *International Journal of Electrical Power & Energy Systems*, vol. 73, pp. 516-527, Jan. 2015.

[51] A. Xue, L. Gu, and H. Hong *et al.*, "Pmu angle deviation detection and correction using line reactive power measurements," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2679-2689, May 2022.

[52] M. Hassouna, C. Holzhüter, and P. Lytaev *et al.*, "Graph reinforcement learning for power grids: a comprehensive survey," *arXiv:2407.04522v1*, 2024.

[53] C. Zhang, Y. Xu, and Z. Y. Dong *et al.*, "Multi-objective adaptive robust voltage/var control for high-PV penetrated distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5288-5300, Nov. 2020.

[54] F. U. Nazir, B. C. Pal, and R. A. Jabr, "Affinely adjustable robust volt/var control without centralized computations," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 656-667, Jan. 2023.

[55] H. T. Nguyen and D.-H. Choi, "Distributionally robust safety filter for learning-based control in active distribution systems," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4972-4975, Nov. 2023.

[56] S. Li, W. Wu, and Y. Lin, "Robust data-driven and fully distributed volt/var control for active distribution networks with multiple virtual power plants," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2627-2638, Jul. 2022.

[57] D. Cao, J. Zhao, and J. Hu *et al.*, "Physics-informed graphical representation-enabled deep reinforcement learning for robust distribution system voltage control," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 233-246, Jan. 2024.

[58] S. Song, H. Xiong, and Y. Lin *et al.*, "Robust three-phase state estimation for pv- integrated unbalanced distribution systems," *Applied Energy*, vol. 322, 2022.

[59] R. Wang, P. Li, and H. Yu *et al.*, "Identification of critical uncertain factors of distribution networks with high penetration of photovoltaics and electric vehicles," *Applied Energy*, vol. 329, 2023.

[60] K. Zhang, T. Sun, and Y. Tao *et al.*, "Robust multi-agent reinforcement learning with model uncertainty," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10571-10583, Nov. 2020.

[61] S. Gu, J. G. Kuba, and M. Wen *et al.*, "Multi-agent constrained policy optimisation," *arXiv:2110.02793*, 2022.

[62] R. Lowe, Y. I. Wu, and A. Tamar *et al.*, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, pp. 10971-10983, Jan. 2017.

[63] H. Zhang, H. Chen, and C. Xiao *et al.*, "Robust deep reinforcement learning against adversarial perturbations on state observations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21024-21037, Nov. 2020.

[64] Z. Liu, Z. Guo, and Z. Cen *et al.*, "On the robustness of safe reinforcement learning under observational perturbations," in *The Eleventh International Conference on Learning Representations*, 2023.

[65] 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.

[66] H. Khodr, F. Olsina, and P. De Oliveira-De Jesus *et al.*, "Maximum savings approach for location and sizing of capacitors in distribution systems," *Electric Power Systems Research*, vol. 78, no. 7, pp. 1192-1203, Jul. 2008.

[67] X. Wu, A. J. Conejo, and N. Amjady, "Robust security constrained acopf via conic programming: Identifying the worst contingencies," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 5884-5891, Nov. 2018.

[68] J. Ackermann, V. Gabler, and T. Osa *et al.*, "Reducing overestimation bias in multi-agent domains using double centralized critics," *arXiv:1910.01465*, 2019.

[69] C. Yu, A. Velu, and E. Vinitsky *et al.*, "The surprising effectiveness of PPO in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24611-24624, May 2022.

[70] Y. Chow, O. Nachum, and E. Duenez-Guzman *et al.*, "A Lyapunov-based approach to safe reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 31, pp. 23501-23514, May 2018.