IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 March 2023, accepted 2 April 2023, date of publication 5 April 2023, date of current version 10 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264687

RESEARCH ARTICLE

Safety-Integrated Online Deep Reinforcement Learning for Mobile Energy Storage System Scheduling and Volt/VAR Control in Power Distribution Networks

SOI JEON[1], (Student Member, IEEE), HOANG TIEN NGUYEN[2], (Graduate Student Member, IEEE), AND DAE-HYUN CHOI[2], (Member, IEEE)

[1]HD Hyundai Global R&D Center, Seongnam-si, Gyeonggi-do 13553, South Korea
[2]School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 156-756, South Korea

Corresponding author: Dae-Hyun Choi (dhchoi@cau.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2022R1F1A1062888, and in part by NRF funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant 2021R1A4A1031019.

ABSTRACT In coupled power distribution and transportation (CPT) system, a joint scheduling framework for mobile energy storage systems (MESSs) and Volt/VAR control (VVC) ensures reliable power distribution grid operations while supporting electric vehicle loads at electric vehicle charging stations (EVCSs). However, conventional model-based optimization methods for MESS scheduling and VVC may yield suboptimal solutions and greater computation times because of MESS operation and VVC in uncertain environment of CPT systems. To resolve this issue, this study proposes a model-free deep reinforcement learning (DRL) framework. In this framework, smart inverters of MESSs and solar photovoltaic (PV) systems cooperate to minimize the real power loss and mitigate the violations of both MESSs' state of charge (SOC) and voltage in the power distribution network, while MESSs travel via the transportation network to satisfy EV loads at EVCSs. A MESS routing algorithm based on Dijkstra's algorithm is developed to determine the optimal destinations of the MESSs. In addition, two safety modules are developed to ensure that neither SOC nor voltage violations occur by adjusting real and/or reactive power of MESSs and PV systems during the training process. The developed MESS routing algorithm and safety modules are integrated into the proposed DRL framework, wherein the DRL agent performs the desired MESS scheduling and VVC through safe exploration during the training procedure. The proposed approach is tested in coupled IEEE 33-bus power distribution and 15-node transportation systems and coupled IEEE 57-bus power distribution and 42-node transportation systems. Numerical examples demonstrate the advantages of the proposed approach in terms of training convergence, real power loss, and SOC/voltage violation.

INDEX TERMS Deep reinforcement learning, safe exploration, mobile energy storage system, Volt/VAR control, smart inverter, coupled power distribution and transportation system.

I. INTRODUCTION

Recent advances in various distributed energy resources (DERs), including solar photovoltaic (PV) systems, energy storage systems (ESSs), and electric vehicles (EVs), have led to an on-going transition from passive to active power distribution networks to ensure efficient and reliable distribution grid operations [1]. However, a rapid fluctuation in PV generation outputs and a sudden increase in aggregated EV loads at EV charging stations (EVCSs) may have detrimental effects

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

on such operations, and cause abnormal voltage regulation with increasing voltage violations, network power losses, and incomplete peak shaving [2].

To mitigate such adverse impacts, much attention has been paid to the utilization of smart inverters of DERs as new voltage regulating devices. Smart inverters rapidly absorb or inject the real/reactive power of DERs from or into the distribution grid to reduce the network power loss [3] and peak load [4] while maintaining a normal voltage level [5] along the active distribution feeder. A real-time voltage control method was proposed in [6] in which the smart inverters of PV systems and ESSs are coordinated for fast voltage regulation in power distribution networks with high PV penetration. In [7], a two-layer Volt/VAR Control (VVC) method using the reactive power voltage adjustment of the PV inverter was presented to achieve the network loss optimization along the computation time reduction. A three-stage robust inverter-based VVC framework was developed in [8] in which the fast voltage regulators (PV system inverters) and slow voltage regulators (on-load tap changers (OLTCs) and capacitor banks (CBs)) cooperate to minimize the total energy loss while addressing the uncertainties of PV outputs and load demands. In [9], a novel algorithm was proposed to determine optimal parameters of volt-var curve for local voltage control using PV inverters to maintain normal voltage quality in active power distribution systems. Recently, mobile energy storage systems (MESSs) have become vital DERs to guarantee stable and efficient active distribution grid operations because of their mobility and plug-and-play functionality at any time and location [10]. In general, MESSs are truck-mounted ESSs owned by a utility and perform the following two tasks by regulating the real and reactive powers of ESSs through their smart inverters: i) service restoration for resilient active distribution grid operations against power failures due to natural disasters and ii) mobile charging for waiting EVs at EVCSs.

Numerous previous studies formulated model-based optimization problems to implement scheduling algorithms for MESSs to achieve the aforementioned two tasks in coupled power and transportation (CPT) networks. A core part of these previous studies was to construct separate and joint constraints of MESS charging/discharging and road routing corresponding to power distribution and transportation networks, respectively. For service restoration, two-stage stochastic optimization models were presented to minimize the total system cost under uncertainties of load consumption, PV generation output, and traffic demand in CPT networks. These models considered the scheduling of MESS fleet, generation dispatching of microgrids and network topology reconfiguration [11] and coordination of MESSs and hybrid AC/DC microgrids [12]. In [13], conventional MESSs were transformed into separable MESSs combined with mobile generators and fuel tanks. A key concept of separable MESSs is that they carry multiple detachable modules, each of which is independently scheduled to perform service restoration along with dynamic network reconfiguration. A tri-level optimization framework was presented in [14], wherein the optimal sizing and pre-positioning problems of MESSs were solved in a decentralized manner to enhance the resilience of networked microgrids. In [15], an optimization-based long-term transmission planning model with stationary ESSs (SESSs) and MESSs was proposed to minimize the investment cost of transmission lines and SESSs/MESSs, operating cost of conventional generators, and transportation cost of MESSs. Concerning MESS-enabled charging of EVs at EVCSs, MESSs were exploited as mobile charging stations (MCSs) dispatched for EV charging at any time and location. These studies aimed to reduce the number of waiting EVs at EVCSs using a new communication scheme [16] while maintaining a normal voltage profile using the reactive power capability of MCSs in the power distribution grid [17]. Additionally, they aimed to relieve the overload of EVCSs with restricted capacity [18], and increase the operational profits of EVCSs through economical and computationally efficient charging of MESSs using a Lyapunov-based method in a distributed manner [19]. In [20], it was shown that MCSs, as opposed to fixed charging stations, can reduce the charging time and cost of EVs owing to their mobility and flexibility. In [21], a self-scheduling model for a smart EVCS integrated with a PV system that combined heat and power, electrical and heat energy storage, and an MCS was presented. This model aimed to maximize the financial gains of smart EVCSs while maintaining adequate peak electrical demand. Its effectiveness was tested on real smart EVCSs located in Los Angeles, California. More recently, a joint optimization framework for VVC and MCS operation was presented in [22]. Herein, voltage regulators, such as OLTCs and CBs, as well as smart inverters of PV systems and MCSs cooperate to realize the following. i) Reduce real power losses, peak demand, and voltage violations, and support EVCS loads in power distribution networks, and ii) minimize the traveling costs of MCSs in transportation networks.

In this study, we assume that MESSs are utilized as MCSs (i.e., the second task of MESSs in the aforementioned literature review) that perform VVC by regulating the real and/or reactive power of MCSs via their smart inverters while supplying the real power of MCSs to EVCS load demands. However, previous studies based on model-based optimization approaches have two major limitations. First, they must rely on inaccurate knowledge of power distribution and transportation systems (e.g., uncertain network topology/line parameters and PV generation outputs in power distribution networks, and dynamically time-varying traffic conditions in transportation networks). Evidently, this leads to incorrect operation schedules for voltage regulators and MCSs. Furthermore, this knowledge varies dynamically and is difficult to obtain in real-world scenarios. Second, as the power distribution and transportation networks become larger, model-based optimization problems become more complex and encompass a larger number of decision variables.

Consequently, they spend notably higher computation time to obtain their optimal solution and even yield infeasible solutions. Therefore, this model-based optimization approach does not scale well for real-time applications.

To resolve these limitations, reinforcement learning (RL) and deep reinforcement learning (DRL) have recently attracted attention as model-free methods for efficient scheduling of EVs/MESSs and VVC. In [23], a model-free RL method using state-action-reward-state-action was employed to jointly determine the pricing and charging scheduling of EVs at EVCSs with random EV arrivals and departures. A model-free DRL approach for real-time scheduling of EV charging and discharging was developed in [24]. This approach was formulated as a Markov decision process (MDP) problem with an unknown transition probability by considering the randomness of the electricity price and commuting behavior of EVs. In [25], a DRL model for energy management of an intelligent solar parking lot (ISPL) was presented. In this model, ISPL calculates both the optimal charging and hydrogen refueling schedules of fuel cell EVs under uncertainties in PV generation output. Additionally, the ISPL estimates the arrival and departure times of EVs, and preferences of EV users. In [26], a DRL-based routing algorithm for multiple EVs mounted with PV panels and ESSs was proposed; herein, EVs were dispatched to supply their power to consumers having uncertain demands. A safe DRL method for the EV charging/discharging problem was developed in [27], wherein the constrained optimal EV charging/discharging schedules are calculated using a deep neural network without defining a penalty term and adjusting its coefficient. In [28], a DRL model employing the deep deterministic policy gradient (DDPG) method was presented to jointly control the room temperature of households and bidirectional EV charging to minimize the electricity cost. A scalable DRL approach for EV routing was proposed in [29], wherein the EV routing problem within a time window was solved on the basis of an attention model incorporating a pointer network and a graph layer to parameterize the stochastic policy of a DRL agent. More recently, a DRL method was utilized to perform the routing and charging/discharging scheduling of MESSs to enhance the resilience of power distribution systems in CPT networks. In [30], a twin-delayed DDPG method was applied to coordinate the scheduling of MESSs and conduct resource dispatching of microgrids for integrated service restoration by considering uncertainties in load consumption. A real-time multi-agent DRL (MADRL) approach using double deep Q-networks was developed for power system service restoration in [31]. In this approach, the structure of the CPT network was explicitly formulated into the environment of DRL agents, wherein the constraints for stable power distribution grid operations and realistic traffic conditions of MESSs were constructed.

In addition, many studies developed DRL-based VVC algorithms using the smart inverters of DERs and EVCSs. A two-stage hybrid VVC framework using MADRL was presented in [32], wherein OLTC and CBs were scheduled using a mixed-integer second-order cone programming method on a slow timescale, and the reactive power of PV systems via their smart inverters was regulated using a DRL method on a fast timescale to mitigate quickly fluctuating voltage violations. A decentralized DRL approach using a soft actor-critic (SAC) was presented in [33], wherein PV inverters, OLTCs, and CBs were coordinated to minimize the voltage violations while reducing the switching frequencies of OLTCs and CBs. In [34], a safe off-policy DRL problem using the constrained SAC (CSAC) method was formulated in a constrained MDP problem to reduce the network power loss and operating cost of voltage regulators while satisfying the operational constraints in power distribution systems. In [35], a model-augmented safe DRL method for VVC was proposed to improve the sampling efficiency and enhance the safety of the DRL agent using a quadratic programming-based policy neural network. A three-stage peak shaving and VVC framework using the online safe DRL method was presented in [36]. In this method, a stand-alone safety module with no modification of existing DRL algorithms was developed to eliminate voltage violations during the training process of the agents. In [37], a MADRL-based VVC framework was developed wherein DRL agents trained for legacy voltage regulators and smart inverters of PV systems cooperated to execute VVC in unbalanced power distribution systems with voltage-dependent loads. A two-stage real-time VVC method using optimization and DRL methods was implemented in power distribution systems with EVs in [38]. In the first stage, the scheduling problem of OLTCs and CBs was formulated as a mixed-integer second-order cone programming problem to minimize power loss. In the second stage, a DRL method using DDPG was employed to build the local voltage control strategy through which both real and reactive power capabilities of EVs at EVCSs are utilized via the smart inverters of the EVCSs to mitigate voltage violations.

However, previous studies on DRL-based MESS scheduling and inverter-based VVC have the following limitations. First, no previous study presented a DRL framework for VVC operation considering joint real and reactive power dispatch of MESS and PV systems via their smart inverters in CPT systems. Obviously, optimal joint MESS scheduling and VVC with MESSs and PV systems would become more difficult under uncertain operation conditions of CPT systems. Therefore, it is necessary to develop a joint DRL framework to address such uncertainties. Second, in [30] and [31], which are similar studies to ours, a DRL model that schedules the operation of MESSs in CPT system was presented. However, in these studies, MESSs were used to enhance the resilience of power distribution grids against extreme natural disasters by scheduling only their real power. Incorrect reactive power dispatch of MESSs may hinder stable power distribution grid operations with high network power loss and abnormal voltage profiles. Third, the road routing process of MESSs in a transportation network is unclear in [30] and [31]. Essentially,

given current locations of MESSs, the method for determining their next locations via optimal road paths was not properly described. Fourth, no safety module for the DRL agent of the MESSs and PV systems was developed in [30] and [31]. Given a prescribed environment, the DRL agent performs a broad exploration during the training period to determine its optimal policy (i.e., the action of the agent). However, such exploration may yield frequent violations of the MESS SOC and voltage limits, and increase the network power loss. Consequently, the performance of the MESSs would be degraded and the operation of the power distribution system operation in real-world scenarios would be destabilized.

To resolve these limitations, we propose a DRL framework that performs VVC by simultaneously controlling the operation of MESSs and PV systems, while calculating the optimal destination of MESSs and guaranteeing the safe operation of MESSs and PV systems in the CPT network. In the present study, we consider MESSs as MCSs that supply real power to EVs and perform voltage regulation at EVCSs. In addition, we consider a situation wherein the agent in the proposed DRL framework directly interacts with the real environment, i.e., power distribution and transportation systems, during its online training process. All in all, the main contributions of this study can be summarized as follows.

- We present a safety-integrated DRL framework using the SAC method. This framework jointly performs real and/or reactive power dispatch of MESSs and PV systems via their smart inverters for VVC to reduce real power loss and support EV loads at EVCSs in CPT networks while maintaining acceptable levels of the SOC of MESSs and voltage magnitude.
- We propose a MESS road routing algorithm based on a Dijkstra's algorithm through which the DRL agent determines the traveling status of MESSs according to its state (i.e., the next arrival time and current location of MESSs) and action (i.e., the destination of MESSs) in the transportation networks.
- We propose two safety modules with a plug-and play functionality for them to be readily integrated into any DRL algorithm. These safety modules enable the DRL agent to perform a safe exploration during the training period by adjusting the real and/or reactive power of MESSs and PV systems, thereby leading to no MESS SOC and voltage violations during both the training and execution periods.
- The performance of the proposed DRL approach is compared with that of conventional mixed-integer linear programming (MILP)-based optimization and DRL approaches without safety module. The simulation results demonstrate the advantages of the proposed approach in terms of real power loss and number of SOC and voltage violations along with the convergence rate of the training curve of the DRL agent.

The remainder of this paper is organized as follows. Section II introduces a CPT system model for active power distribution system operation and MESS scheduling along with the architecture of the proposed DRL approach. Section III formulates a safety-integrated DRL algorithm for joint scheduling of MESS and PV system operations for VVC. Section IV presents the simulation results for the proposed algorithm. Finally, Section V concludes the study.

II. SYSTEM MODEL

Let us consider a CPT system wherein MESSs and PV systems are coordinated to minimize the real power loss and voltage violations for VVC in active power distribution systems while MESSs support EV loads at EVCSs. In this study, MESSs are assumed to be truck-mounted EVs having an ESS. They are dispatched through a transportation network to the EVCSs to perform the following two tasks using their smart inverters. i) Real power charging from the grid and real power discharging to EVs via EVCSs that are located at the intersection of the power distribution and transportation networks; ii) reactive power absorption and injection from and to the grid for voltage regulation. VVC aims to conduct voltage regulation via the smart inverters of MESSs and PV systems using their reactive power. MESSs, PV systems, and EVCS/non-EVCS loads are connected to an electric bus $b \in \mathcal{B}$ in the active power distribution networks wherein the MESSs and PV systems belong to the subsets $\mathcal{B}^{\text{MESS}}$ and $\mathcal{B}^{\text{PV}}$ of the set $\mathcal{B}$, respectively. Each MESS $m \in \mathcal{M}$ departs from its depot, transits via transportation node $i \in \mathcal{I}$, and arrives at the EVCSs to support the EV loads and perform voltage regulation. The transportation node consists of a node with an EVCS ($i \in \mathcal{I}^{\text{EVCS}}$) and a node without an EVCS ($i \in \mathcal{I}^{\text{non-EVCS}}$). Node $i \in \mathcal{I}^{\text{EVCS}}$ can be the same as bus $\mathcal{B}^{\text{MESS}}$ owing to the intersection of the power distribution and transportation networks. $|\mathcal{A}|$ represents the cardinality of set $\mathcal{A}$. The bold font denotes a vector. The detailed operational constraints of MESSs and PV systems in the CPT system are described in the following subsections.

A. ACTIVE POWER DISTRIBUTION GRID OPERATION

DistFlow equations [39] of the linearized real ($P^{\text{line}}_{hb,t}$) and reactive ($Q^{\text{line}}_{hb,t}$) power flows at line hb and the squared voltage magnitude ($v_{b,t} = (V_{b,t})^2$) at bus b and time t are written as follows.

$$P^{\text{line}}_{hb,t} = \sum_{k \in \mathcal{B}_b} P^{\text{line}}_{bk,t} + P^{\text{node}}_{b,t} \tag{1}$$

$$Q^{\text{line}}_{hb,t} = \sum_{k \in \mathcal{B}_b} Q^{\text{line}}_{bk,t} + Q^{\text{node}}_{b,t} \tag{2}$$

$$v_{b,t} = v_{h,t} - 2(r_{hb}P^{\text{line}}_{hb,t} + x_{hb}Q^{\text{line}}_{hb,t}). \tag{3}$$

Equations (1) and (2) indicate the real and reactive power flow balance at bus b and time t, where bus k is included in a set of buses $\mathcal{B}_b$ having all neighboring buses to bus b. The squared voltage drop between buses h and b is described in (3) where the resistance and reactance at line hb are denoted by r_{hb} and x_{hb}, respectively.

Equations (4) and (5) represent the net real ($P^{\text{node}}_{b,t}$) and reactive ($Q^{\text{node}}_{b,t}$) power consumptions belonging

to (1) and (2). They are written in terms of real/reactive power consumption ($P^{\text{load}}_{b,t}$, $Q^{\text{load}}_{b,t}$), real/reactive power generation of the PV system ($\widehat{P}^{\text{PV}}_{b,t}$, $Q^{\text{PV}}_{b,t}$), and real/reactive charging/discharging power of MESS m at bus b and time t ($P^{\text{MESS,ch}}_{b,t,m}$, $P^{\text{MESS,dch}}_{b,t,m}$, $Q^{\text{MESS}}_{b,t,m}$).

$$P^{\text{node}}_{b,t} = P^{\text{load}}_{b,t} - \widehat{P}^{\text{PV}}_{b,t} + \sum_{m\in\mathcal{M}} \left(P^{\text{MESS,ch}}_{b,m,t} - P^{\text{MESS,dch}}_{b,m,t}\right) \quad (4)$$

$$Q^{\text{node}}_{b,t} = Q^{\text{load}}_{b,t} + Q^{\text{PV}}_{b,t} + \sum_{m\in\mathcal{M}} Q^{\text{MESS}}_{b,m,t}. \quad (5)$$

The net real/reactive power consumption ($P^{\text{load}}_{b,t}$, $Q^{\text{load}}_{b,t}$) in (4) and (5) are expressed in terms of ZIP load models, which are defined in (6) and (7):

$$P^{\text{load}}_{b,t} = \widehat{P}^{\text{load,nom}}_{b,t}\left[Z_p\left(\frac{V_{b,t}}{V_0}\right)^2 + I_p\left(\frac{V_{b,t}}{V_0}\right) + P_p\right] \quad (6)$$

$$Q^{\text{load}}_{b,t} = \widehat{Q}^{\text{load,nom}}_{b,t}\left[Z_q\left(\frac{V_{b,t}}{V_0}\right)^2 + I_q\left(\frac{V_{b,t}}{V_0}\right) + P_q\right] \quad (7)$$

where $\widehat{P}(Q)^{\text{load,nom}}_{b,t}$ is the predicted real (reactive) power consumption at nominal voltage (V_0). The percentage of constant impedance/current/power load for real (reactive) power follows its coefficient $\{Z_{p(q)}, I_{p(q)}, P_{p(q)}\}$ with $Z_{p(q)} + I_{p(q)} + P_{p(q)} = 1$.

Equations (8) and (9) represent the limits of the charging (discharging) real power ($P^{\text{MESS,ch(dch)}}_{b,m,t}$) and reactive power ($Q^{\text{MESS}}_{b,m,t}$) of the MESS, respectively, where $b^{\text{ch(dch)}}_{b,m,t}$ and $S^{\text{MESS}}_{b,m}$ denote the binary charging (discharging) decision variable and apparent power of MESS m at bus b, respectively. The reactive power of the PV system is constrained by the apparent power (S^{PV}_b) and predicted output ($\widehat{P}^{\text{PV}}_{b,t}$) of the PV system at bus b and time t, as described in (10). The squared voltage magnitude $v_{b,t}$ at bus b and time t is limited by $v^{\min} = 0.95^2$ and $v^{\max} = 1.05^2$ according to (11).

$$0 \le P^{\text{MESS,ch(dch)}}_{b,m,t} \le b^{\text{ch(dch)}}_{b,m,t} P^{\text{MESS,ch(dch),max}} \quad (8)$$

$$\left(P^{\text{MESS,ch}}_{b,m,t} + P^{\text{MESS,dch}}_{b,m,t}\right)^2 + \left(Q^{\text{MESS}}_{b,m,t}\right)^2 \le \left(S^{\text{MESS}}_{b,m}\right)^2 \quad (9)$$

$$\left(\widehat{P}^{\text{PV}}_{b,t}\right)^2 + \left(Q^{\text{PV}}_{b,t}\right)^2 \le \left(S^{\text{PV}}_b\right)^2 \quad (10)$$

$$v^{\min} \le v_{b,t} \le v^{\max}. \quad (11)$$

B. MESS OPERATION IN THE CPT NETWORK

Equation (12) indicates that any MESS m arrives at one transportation node at most at time t, with $b^{\text{a}}_{i,m,t}$ denoting a binary decision variable that determines the arrival status of the MESS at node i. Given that there is no MESS travel (i.e., $b^{\text{tr}}_{m,t} = 0$), charging or discharging the MESS can be carried out according to (13). Equation (14) allows the MESS to charge or discharge only when it arrives at the EVCS (i.e., $b^{\text{a}}_{i,m,t} = 1$).

$$\sum_{i\in\mathcal{I}} b^{\text{a}}_{i,m,t} \le 1 \quad (12)$$

$$\sum_{i\in\mathcal{I}} b^{\text{ch(dch)}}_{i,m,t} \le 1 - b^{\text{tr}}_{m,t} \quad (13)$$

$$b^{\text{ch}}_{i,m,t} + b^{\text{dch}}_{i,m,t} \le b^{\text{a}}_{i,m,t}. \quad (14)$$

Equation (15) describes the change in the SOC of MESS m at time t with the scheduling time unit Δt. This is expressed using the SOC at previous time $t-1$, sum of charging/discharging real power ($\sum_{b\in\mathcal{B}^{\text{MESS}}} P^{\text{MESS,ch}}_{b,m,t}$, $\sum_{b\in\mathcal{B}^{\text{MESS}}} P^{\text{MESS,dch}}_{b,m,t}$) at all EVCSs, charging/discharging efficiency (η^{ch}_m, η^{dch}_m), battery capacity ($E^{\text{MESS,max}}_m$), binary traveling status ($b^{\text{tr}}_{m,t}$), and traveling efficiency (η^{tr}_m [kWh/Δt]). Equations (16) and (17) limits the SOC at time $t \ne T$ and time $t = T$, respectively, given scheduling period $t \in \mathcal{T} = \{1, \ldots, T\}$. Here, T is the finishing scheduling period of MESSs. Note that the condition expressed by (17) enables the MESS to be operated correctly the next day while maintaining the desired SOC level, i.e., $SOC^{\text{r}}_{m,T}$ of the MESS m at $t = T$.

$$SOC_{m,t} = SOC_{m,t-1} + \left(\eta^{\text{ch}}_m \sum_{b\in\mathcal{B}^{\text{MESS}}} P^{\text{MESS,ch}}_{b,m,t} - \eta^{\text{dch}}_m \sum_{b\in\mathcal{B}^{\text{MESS}}} P^{\text{MESS,dch}}_{b,m,t} - \eta^{\text{tr}}_m b^{\text{tr}}_{m,t}\right)\Delta t / E^{\text{MESS,max}}_m \quad (15)$$

$$SOC^{\min}_m \le SOC_{m,t} \le SOC^{\max}_m \quad (16)$$

$$SOC^{\text{r}}_{m,T} \le SOC_{m,T}. \quad (17)$$

C. MESS ROAD ROUTING IN THE TRANSPORTATION NETWORK

According to (18), each MESS m builds a single transit path at most in the scheduling time, where $b^{\text{c}}_{ij,m,t}$ determines the binary connection status of the transit path $i-j$. Equation (19) enforces that the traveling path $i-j$ is connected ($b^{\text{c}}_{ij,m,t} = 1$) when the MESS arrives at the EVCS ($b^{\text{a}}_{i,m,t} = 1$). Equation (20) guarantees that during the normalized traveling time $\gamma_{ij,t}$, no MESS is allowed to stay at any node when the transit path $i-j$ is constructed. Here, $\gamma_{ij,t}$ represents the normalized traveling time for path $i-j$ and is defined as $\gamma_{ij,t} = \lceil \frac{w_{ij,t}}{\Delta t} \rceil$. Herein, $w_{ij,t}$ is the ratio of the distance (d_{ij}) of transit path $i-j$, which is obtained by Dijkstra's algorithm, to the average driving speed of any MESS, Δt is the scheduling time unit, and $\lceil\cdot\rceil$ is a round-up. According to (21), the MESS is forced to arrive at node j after $\gamma_{ij,t}$ via transit path $i-j$. Equation (22) ensures that no transit path is constructed while the MESS remains at the EVCS node. According to (23) no MESS travel is allowed ($b^{\text{tr}}_{m,t} = 0$) when the MESS arrives at node i

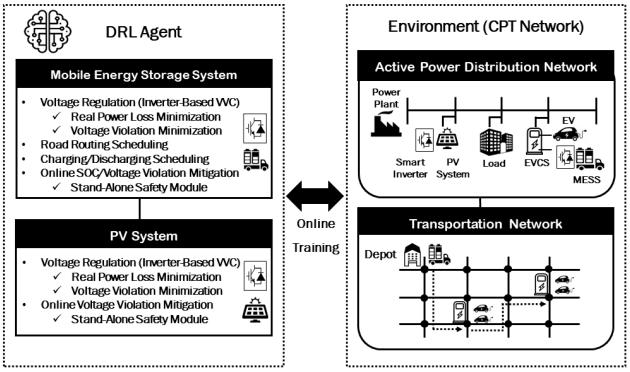


FIGURE 1. Architecture of the proposed DRL framework in the CPT network.

($b^{\mathrm{a}}_{i,m,t} = 1$) with no connected traveling path ($b^{\mathrm{c}}_{ij,m,t} = 0$).

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}_i} b^{\mathrm{c}}_{ij,m,t} \le 1 \tag{18}$$

$$b^{\mathrm{a}}_{i,m,t} - 1 \le \sum_{j\in\mathcal{I}_i} b^{\mathrm{c}}_{ij,m,t} \le b^{\mathrm{a}}_{i,m,t} \tag{19}$$

$$b^{\mathrm{c}}_{ij,m,t} + \frac{1}{|\mathcal{I}||\Gamma|}\sum_{t'=t+1}^{t+\gamma_{ij,t}-1}\sum_{j\neq i\in\mathcal{I}} b^{\mathrm{a}}_{j,m,t'} \le 1 \tag{20}$$

$$b^{\mathrm{c}}_{ij,m,t} \le b^{\mathrm{a}}_{j,t+\gamma_{ij,t},m} \tag{21}$$

$$b^{\mathrm{a}}_{i,m,t} - \sum_{j\in\mathcal{I}_i} b^{\mathrm{c}}_{ij,m,t} \le b^{\mathrm{a}}_{i,m,t+1} \tag{22}$$

$$b^{\mathrm{tr}}_{t,m} = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}_i} b^{\mathrm{c}}_{ij,m,t} - \sum_{i\in\mathcal{I}} b^{\mathrm{a}}_{i,m,t} + 1. \tag{23}$$

D. ARCHITECTURE OF THE PROPOSED DRL FRAMEWORK

As shown in Fig. 1, we present a DRL framework wherein a DRL agent performs VVC and satisfies the EV loads at EVCSs by jointly scheduling the operation of MESSs and PV systems while interacting with a real-world CPT environment. The DRL agent enables the MESSs to conduct the following three tasks. i) Scheduling the road routing of multiple MESSs in the transportation network where the destinations of the MESSs (i.e., the locations of EVCSs) are determined upon their current locations and traveling times calculated by Dijkstra's algorithm [40]; ii) scheduling the real charging/discharging power of MESSs from/to the grid and real discharging power to the EVs at EVCSs via their smart inverters in the power distribution network; and iii) scheduling the reactive charging and discharging power of MESSs from and to the grid for VVC (i.e., minimizing real power loss and voltage violations) via their smart inverters in the power distribution network. Additionally, the DRL agent conducts a VVC by scheduling the reactive charging and discharging power dispatch of PV systems via their smart inverters. The DRL agent is equipped with stand-alone safety modules that enable the MESSs and PV systems to mitigate the SOC violations of the MESSs and nodal voltage violations during the online training process. In summary, the proposed safety-integrated DRL framework performs VVC and supports EV loads at EVCSs by scheduling the operation of MESSs and PV systems while significantly reducing SOC violations of MESSs and nodal voltage violations during both the training and execution stages of the DRL process.

III. SAFE DRL-BASED MESS SCHEDULING AND VVC ALGORITHM

In this section, we formulate a safety-integrated DRL algorithm for joint MESS and PV system scheduling to perform VVC considering a safe exploration of the DRL agent during its online training process. Sections III-A and III-B provide an overview of the background of MDP, RL, and SAC, which is a state-of-the-art DRL method. The state/action space and reward function for the DRL agent are formulated using the SAC method in Section III-C. Section III-D presents a MESS routing algorithm based on Dijkstra's algorithm that updates the states of the MESS (i.e., the current location, arrival time at the destination, and transit status of the MESS) based on its action defined in Section III-C. Sections III-E and III-F propose two safety algorithms that prevent both SOC violations of MESSs and nodal voltage violations during the training process by adjusting MESS real charging/discharging power and MESS/PV reactive power, respectively. Given that the charging and discharging of MESS influence the voltage level, the safety algorithm presented in Section III-E is performed prior to the execution of the safety algorithm reported in Section III-F.

A. MDP AND RL

An MDP is an environment wherein RL methods can be mathematically formulated. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R)$, wherein $\mathcal{S}$ and $\mathcal{A}$ represent the sets of states s_t and actions a_t at current time t for an RL agent, respectively. $P : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to [0,\infty)$ denotes the transition probability from the current state $s_t \in \mathcal{S}$ to the next state $s_{t+1} \in \mathcal{S}$ via action $a_t \in \mathcal{A}$. $R : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ represents a numerical reward for the agent transition, where r_t is formulated as $r_t = R(s_t, a_t, s_{t+1})$. Each transition in the MDP holds the Markov property (i.e., the state transition relies only on the current state). The RL agent aims to find an optimal policy $\pi^\star$ that maximizes the discounted cumulative future rewards $J(\pi) = \mathbb{E}\left[\sum_{t=0}^{N_T}\gamma^t r_t\right]$, where policy π is the probability distribution over the actions under each state $s_t \in \mathcal{S}$, (i.e., $a_t \sim \pi(\cdot|s_t)$), $\gamma \in [0,1)$ is the discount rate that indicates the relative importance of the future reward to the current reward, and N_T is the terminal time of the agent's learning process. Concerning the RL method, the state-value function $V_\pi(s)$ and state-action value function $Q_\pi(s,a)$ are, respectively, formulated as follows.

$$V_\pi(s) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{N_T}\gamma^t r_t | s_0 = s\right], \tag{24}$$

$$Q_\pi(s,a) = \mathop{\mathbb{E}}_{\tau\sim\pi}\left[\sum_{t=0}^{N_T}\gamma^t r_t | s_0 = s, a_0 = a\right] \tag{25}$$

where τ represents a trajectory of state and action of the agent under policy π.

B. SAC METHOD

SAC [41] is a state-of-the-art off-policy DRL method that enhances the sample efficiency and training stability of a DRL agent, given continuous state and action spaces. Traditional on-policy DRL approaches need a new sample from the environment to build a gradient step. This leads to poor sample efficiency, thereby preventing agents from performing online training in actual power distribution grids. However, the off-policy-based SAC approach learns the optimal policy by maximizing the following reward function with an additional entropy term: $H(\pi(\cdot|s_t))$: $J(\pi) = \mathbb{E}\left[\sum_{t=0}^{N_T} \gamma^t (r_t + \alpha H(\pi(\cdot|s_t)))\right]$. Here, α represents the temperature coefficient that determines the relative importance of the reward and entropy. The entropy term enables the agent to perform a wider exploration during the learning procedure and improves the sample efficiency of the agent. To ensure the stable convergence of the training curves of the DRL agents using the SAC method, an experience replay method is employed such that the agent's experience at each time step is stored in a replay buffer.

The SAC-based DRL agent is built on five neural networks: a critic network comprising two Q-networks, a value network, a target network, and an actor network. The two Q-networks are denoted by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$; the value and target value networks are denoted by $\boldsymbol{\psi}$ and $\boldsymbol{\psi}_{\text{targ}}$, respectively; and the actor network is denoted by $\boldsymbol{\phi}$. The value, critic, and actor networks are updated by minimizing the following three loss functions, respectively:

$$J_V(\boldsymbol{\psi}) = \mathop{\mathbb{E}}_{(s_t,a_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(V_{\boldsymbol{\psi}}(s_t) - \mathop{\mathbb{E}}_{a_t\sim\pi_{\boldsymbol{\phi}}}\left[\min_{i=1,2} Q_{\boldsymbol{\theta}_i}(s_t,a_t) - \log\pi_{\boldsymbol{\phi}}(a_t|s_t)\right]\right)^2\right] \tag{26}$$

$$J_Q(\boldsymbol{\theta}_i) = \mathop{\mathbb{E}}_{(s_t,a_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_{\boldsymbol{\theta}_i}(s_t,a_t) - \left(r_t + \gamma \mathop{\mathbb{E}}_{s_{t+1}\sim P}\left[V_{\boldsymbol{\psi}_{\text{targ}}}(s_{t+1})\right]\right)\right)^2\right] \tag{27}$$

$$J_\pi(\boldsymbol{\phi}) = \mathop{\mathbb{E}}_{\substack{s_t\sim\mathcal{D}\\ \xi_t\sim\mathcal{N}}}\left[\log\pi_{\boldsymbol{\phi}}(a_t|s_t) - \min_{i=1,2} Q_{\boldsymbol{\theta}_i}(s_t,a_t)\right] \tag{28}$$

where actions are sampled from a standard normal distribution ξ_t (i.e., $a_t = \tanh\left(\mu_{\boldsymbol{\phi}}(s_t) + \sigma_{\boldsymbol{\phi}}(s_t)\odot\xi_t\right)$, $\xi_t \sim \mathcal{N}(0,\mathbf{I})$). In this action sampling, $\mu_{\boldsymbol{\phi}}$ and $\sigma_{\boldsymbol{\phi}}$ represent the mean and log standard deviation, respectively, and they are two outputs of the actor network; $\odot$ denotes the dot product. In addition, a delayed update of the value function is employed to enhance the training stability of the SAC agent, wherein the target value network is updated by $\boldsymbol{\psi}_{\text{targ}} = \tau\boldsymbol{\psi} + (1-\tau)\boldsymbol{\psi}_{\text{targ}}$, where τ is in the range [0,1].

C. MATHEMATICAL FORMULATION FOR DRL AGENT

1) STATE

For each scheduling period $t \in \mathcal{T} = \{1,\ldots,T\}$ based on a 15-min resolution, the state of the DRL agent, denoted by $\mathbf{s}_t$, is described as follows.

$$\mathbf{s}_t = \{t, \mathbf{k}_t^{\text{a}}, \mathbf{l}_t^{\text{c}}, \mathbf{b}_t^{\text{tr}}, \mathbf{SOC}_t, \mathbf{P}_t^{\text{load}}, \mathbf{Q}_t^{\text{load}}, \widehat{\mathbf{P}}_t^{\text{PV}}\}. \tag{29}$$

In (29), the vector of arrival times of MESSs at EVCSs after time t is denoted by $\mathbf{k}_t^{\text{a}} = [k_{1,t}^{\text{a}},\ldots,k_{|\mathcal{M}|,t}^{\text{a}}]$. The vector of current locations of MESSs at time t is defined as $\mathbf{l}_t^{\text{c}} = [l_{1,t}^{\text{c}},\ldots,l_{|\mathcal{M}|,t}^{\text{c}}]$, wherein $l_{m,t}^{\text{c}}$ includes both the location of EVCS $i_{m,t} \in \mathcal{I}^{\text{EVCS}}$ and road path $r(i_{m,t},j_{m,t}) \in \mathcal{R}^{\text{EVCS}}$ between EVCSs $i_{m,t}$ and $j_{m,t}$ for MESS m at time t. Note that the arrival time $k_{m,t}^{\text{a}}$ and road path $r(i_{m,t},j_{m,t})$ of MESS m at time t are determined by Dijkstra's algorithm. The vector of transit statuses of MESSs at time t is denoted by $\mathbf{b}_t^{\text{tr}} = [b_{1,t}^{\text{tr}},\ldots,b_{|\mathcal{M}|,t}^{\text{tr}}]$, wherein $b_{m,t}$ represent the binary transit status of MESS m at time t and $b_{m,t} = 1$ on the transit of MESS m at time t; otherwise, $b_{m,t} = 0$. The vector of SOCs for the MESSs at time t is denoted by $\mathbf{SOC}_t = [SOC_{1,t},\ldots,SOC_{|\mathcal{M}|,t}]$. The vector of nodal real (reactive) powers of the loads at time t is defined as $\mathbf{P}(\mathbf{Q})_t^{\text{load}} = [P(Q)_{1,t},\ldots,P(Q)_{|\mathcal{B}|,t}]$. The vector of the predicted PV real power generation outputs is denoted by $\widehat{\mathbf{P}}_t^{\text{PV}} = [\widehat{P}_{1,t}^{\text{PV}},\ldots,\widehat{P}_{|\mathcal{B}^{\text{PV}}|,t}^{\text{PV}}]$.

2) ACTION

The DRL agent computes three types of actions; these actions are associated with i) road routing (i.e., determination of the destination with EVCS for MESSs), ii) real/reactive charging/discharging power dispatch of MESSs, and iii) reactive power dispatch of PV systems. The action $\mathbf{a}_t$ of the DRL agent at time t is defined as follows.

$$\mathbf{a}_t = \{\mathbf{l}_t^{\text{d}}, \boldsymbol{\alpha}_t^{\text{P,MESS}}, \boldsymbol{\alpha}_t^{\text{Q,MESS}}, \boldsymbol{\alpha}_t^{\text{Q,PV}}\}. \tag{30}$$

In (30), vector $\mathbf{l}_t^{\text{d}} = [l_{1,t}^{\text{d}},\ldots,l_{|\mathcal{M}|,t}^{\text{d}}]$ represents the scheduled destinations for MESSs at time t. Vectors $\boldsymbol{\alpha}_t^{\text{P,MESS}} = [\alpha_{1,1,t}^{\text{P,MESS}},\ldots,\alpha_{|\mathcal{B}^{\text{MESS}}|,|\mathcal{M}|,t}^{\text{P,MESS}}]$, $\boldsymbol{\alpha}_t^{\text{Q,MESS}} = [\alpha_{1,1,t}^{\text{Q,MESS}},\ldots,\alpha_{|\mathcal{B}^{\text{MESS}}|,|\mathcal{M}|,t}^{\text{Q,MESS}}]$, and $\boldsymbol{\alpha}_t^{\text{Q,PV}} = [\alpha_{1,t}^{\text{Q,PV}},\ldots,\alpha_{|\mathcal{B}^{\text{PV}}|,t}^{\text{Q,PV}}]$ determine the dispatched real/reactive power of MESSs and reactive power of PV systems at time t, respectively, where $-1 \le \alpha_{b,m,t}^{\text{P,MESS}}, \alpha_{b,m,t}^{\text{Q,MESS}}, \alpha_{b,t}^{\text{Q,PV}} \le 1$. Based on the values of $\boldsymbol{\alpha}_t^{\text{P,MESS}}$, $\boldsymbol{\alpha}_t^{\text{Q,MESS}}$, and $\boldsymbol{\alpha}_t^{\text{Q,PV}}$, the real and reactive powers of the MESSs and PV systems are calculated as follows.

$$P_{b,m,t}^{\text{agent}} = \alpha_{b,m,t}^{\text{P,MESS}}(1-b_{m,t}^{\text{tr}})P_m^{\text{MESS,max}} \tag{31}$$

$$Q_{b,m,t}^{\text{agent}} = \alpha_{b,m,t}^{\text{Q,MESS}}(1-b_{m,t}^{\text{tr}})\sqrt{(S_m)^2 - \left(P_{b,m,t}\right)^2} \tag{32}$$

$$Q_{b,t}^{\text{agent}} = \alpha_{b,t}^{\text{Q,PV}}\sqrt{(S_b)^2 - \left(\widehat{P}_{b,t}\right)^2} \tag{33}$$

where $P_m^{\text{MESS,max}}$ is the maximum limit of the charging/discharging real power of MESS m; S_m and S_b are the apparent powers of MESS m and the PV system at bus b, respectively. Note that from (31) and (32), the real and reactive

charging/discharging processes of MESSs are allowed only when they remain at bus b (i.e., $b^{\mathrm{tr}}_{m,t} = 0$).

3) REWARD FUNCTION

The total reward function r_t of the DRL agent at time t is formulated as a weighted multi-negative cost function that comprises the following five terms.

$$\begin{aligned} r_t = &- \omega_1 \sum_{hb\in\mathcal{L}} P^{\mathrm{loss}}_{hb,t} - \omega_2 \sum_{b\in\mathcal{B}^{\mathrm{PV}}} |\Delta Q_{b,t}| \\ &- \omega_3 \sum_{b\in\mathcal{B}^{\mathrm{MESS}}} \sum_{m\in\mathcal{M}} |\Delta Q_{b,m,t}| \\ &- \omega_4 \sum_{m\in\mathcal{M}} |\Delta SOC_{m,t}| - \omega_5 \sum_{m\in\mathcal{M}} \Delta SOC^{\mathrm{r}}_{m,t} \end{aligned} \tag{34}$$

where

$$P^{\mathrm{loss}}_{hb,t} = \frac{r_{hb}\left[(P^{\mathrm{line}}_{hb,t})^2 + (Q^{\mathrm{line}}_{hb,t})^2\right]}{(V_0)^2}, \quad \forall hb \in \mathcal{L} \tag{35}$$

$$\Delta Q_{b,t} = Q^{\mathrm{agent}}_{b,t} - Q_{b,t}, \quad \forall b \in \mathcal{B}^{\mathrm{PV}} \tag{36}$$

$$\Delta Q_{b,m,t} = Q^{\mathrm{agent}}_{b,m,t} - Q_{b,m,t}, \quad \forall b \in \mathcal{B}^{\mathrm{MESS}},\ m \in \mathcal{M} \tag{37}$$

$$\Delta SOC_{m,t} = SOC^{\mathrm{agent}}_{m,t} - SOC_{m,t}, \quad \forall m \in \mathcal{M} \tag{38}$$

$$\Delta SOC^{\mathrm{r}}_{m,T} = \max(SOC^{\mathrm{r}}_{m,T} - SOC_{m,T}, 0), \quad \forall m \in \mathcal{M}. \tag{39}$$

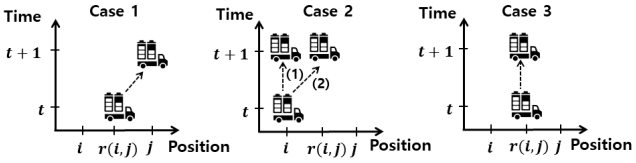
Concerning the five cost functions in (34), the first term is the total real power loss ($P^{\mathrm{loss}}_{hb,t}$) for all the distribution lines. This term is formulated in (35). The second and third terms are the total reactive power mismatches between the actions ($Q^{\mathrm{agent}}_{b,m,t}$, $Q^{\mathrm{agent}}_{b,t}$) and the actual reactive power outputs ($Q_{b,m,t}$, $Q_{b,t}$) of MESS m and PV system at bus b and time t, respectively. These terms are defined in (36) and (37). The non-zero reactive power mismatch in (36) and (37) implies that some voltage violations occur in power distribution systems when actions are applied. The fourth term is the total SOC mismatch between the action ($SOC^{\mathrm{agent}}_{m,t}$) and actual SOC ($SOC_{m,t}$) of MESS m at time t; this term is defined in (38). Note that the mismatches of the reactive power and SOC in the second, third, and fourth terms are significantly reduced by the safety modules of the MESSs and PV systems, which is explained in Sections III-E and III-F. The fifth term is the total positive SOC deviation of the SOC ($SOC_{m,T}$) of MESS m at time T from the required SOC ($SOC^{\mathrm{r}}_{m,T}$); this term is defined in (39). Here, $SOC^{\mathrm{r}}_{m,T}$ guarantees that the MESS successfully returns to the depot and performs its task the following day.

D. ROAD ROUTING ALGORITHM OF MESS

In the proposed approach, the traveling statuses of MESSs in the transportation network are determined by road routing algorithm based on Dijkstra algorithm, as shown in Algorithm 1. As a preliminary step, each MESS m obtains three states (arrival time $k^{\mathrm{a}}_{m,t}$ at the destination after time t, current location $l^{\mathrm{c}}_{m,t}$ and transit status $b^{\mathrm{tr}}_{m,t}$ at time t) and the action $l^{\mathrm{d}}_{m,t}$ of the destination at time t. Given these states and action, Algorithm 1 calculates the MESS's next states at time $t+1$ (i.e., $b^{\mathrm{tr}}_{m,t+1}$, $k^{\mathrm{a}}_{m,t+1}$, and $l^{\mathrm{c}}_{m,t+1}$).

Algorithm 1 MESS Routing in Transportation Network Based on Dijkstra's Algorithm

Take the next arrival time $\mathbf{k}^{\mathrm{a}}_t = [k^{\mathrm{a}}_{1,t}, \ldots, k^{\mathrm{a}}_{|\mathcal{M}|,t}]$, current location $\mathbf{l}^{\mathrm{c}}_t = [l^{\mathrm{c}}_{1,t}, \ldots, l^{\mathrm{c}}_{|\mathcal{M}|,t}]$, and the transit status $\mathbf{b}^{\mathrm{tr}}_t = [b^{\mathrm{tr}}_{1,t}, \ldots, b^{\mathrm{tr}}_{|\mathcal{M}|,t}]$ of MESSs in $\mathcal{S}_t$;
Take the destination $\mathbf{l}^{\mathrm{d}}_t = [l^{\mathrm{d}}_{1,t}, \ldots, l^{\mathrm{d}}_{|\mathcal{M}|,t}]$ of MESSs in $\mathcal{A}_t$;
foreach *MESS m* **do**
  if $t = k^{a}_{m,t} - 1$ **then** *(%Case 1)*
    $b^{\mathrm{tr}}_{m,t+1} = 0$, $k^{\mathrm{a}}_{m,t+1} = k^{\mathrm{a}}_{m,t}$, $l^{\mathrm{c}}_{m,t+1} = j^{\mathrm{a}}_{m,t}$, $j^{\mathrm{a}}_{m,t} \in \mathcal{I}^{\mathrm{EVCS}}$;
  else if $t = k^{a}_{m,t}$ **then** *(%Case 2)*
    if $l^{c}_{m,t} = l^{d}_{m,t}$ **then** *(%Case 2-1)*
      $b^{\mathrm{tr}}_{m,t+1} = 0$, $k^{\mathrm{a}}_{m,t+1} = t+1$, $l^{\mathrm{c}}_{m,t+1} = l^{\mathrm{c}}_{m,t}$, $l^{\mathrm{c}}_{m,t}, l^{\mathrm{c}}_{m,t+1} \in \mathcal{I}^{\mathrm{EVCS}}$;
    else *(%Case 2-2)*
      $b^{\mathrm{tr}}_{m,t+1} = 1$, $k^{\mathrm{a}}_{m,t+1} = t + \gamma_{l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t}} + 1$, $l^{\mathrm{c}}_{m,t+1} = r(l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t})$, $l^{\mathrm{c}}_{m,t+1} \in \mathcal{R}^{\mathrm{EVCS}}$;
    end
  else *(%Case 3)*
    $b^{\mathrm{tr}}_{m,t+1} = 1$, $k^{\mathrm{a}}_{m,t+1} = k^{\mathrm{a}}_{m,t}$, $l^{\mathrm{c}}_{m,t+1} = l^{\mathrm{c}}_{m,t}$, $l^{\mathrm{c}}_{m,t}, l^{\mathrm{c}}_{m,t+1} \in \mathcal{R}^{\mathrm{EVCS}}$;
  end
end



Status	$l^{\mathrm{d}}_{m,t}$	$l^{\mathrm{c}}_{m,t}$	$b^{\mathrm{tr}}_{m,t}$	$l^{\mathrm{c}}_{m,t+1}$	$b^{\mathrm{tr}}_{m,t+1}$
Case 1	-	$r(i,j)$	1	j	0
Case 2-1	i	i	0	i	0
Case 2-2	j	i	0	$r(i,j)$	1
Case 3	-	$r(i,j)$	1	$r(i,j)$	1

FIGURE 2. Illustrative example of MESS road routing using Algorithm 1.

Algorithm 1 comprises three cases: when i) MESS is on the transit one scheduling time unit before arrival at the destination (Case 1), ii) MESS arrives at the destination (Case 2), and iii) MESS is on the transit two additional scheduling time units before arrival at the destination (Case 3). In Case 1, the MESS travels at time t ($b^{\mathrm{tr}}_{m,t} = 1$). Given that it arrives at the destination at time $t+1 = k^{\mathrm{a}}_{m,t}$, the transit status, arrival time, and location of the MESS at time $t+1$ are updated by $b^{\mathrm{tr}}_{m,t+1} = 0$, $k^{\mathrm{a}}_{m,t+1} = k^{\mathrm{a}}_{m,t}$, and $l^{\mathrm{c}}_{m,t+1} = j^{\mathrm{a}}_{m,t}$, respectively. Here, $j^{\mathrm{a}}_{m,t}$ is the index of the next arrival of EVCS when the MESS departs from current location $l^{\mathrm{c}}_{m,t} = r(i_{m,t}, j_{m,t})$, which is determined by the agent's action prior to traveling. In Case 2, the MESS arrives at its destination at $t = k^{\mathrm{a}}_{m,t}$.

Case 2 is decomposed into two subcases (Cases 2-1 and 2-2) according to the action of the MESS, i.e., $l^{\mathrm{d}}_{m,t}$. If the MESS wishes to stay and perform the charging and discharging processes at the next time (i.e., $l^{\mathrm{c}}_{m,t} = l^{\mathrm{d}}_{m,t}$) (Case 2-1), the MESS remains at $t+1$ ($b^{\mathrm{tr}}_{m,t+1} = 0, l^{\mathrm{c}}_{m,t+1} = l^{\mathrm{c}}_{m,t}$) and the arrival time is updated by $k^{\mathrm{a}}_{m,t+1} = t+1$. Otherwise (Case 2-2), the MESS leaves for another destination determined by the new action $l^{\mathrm{d}}_{m,t}$ ($b^{\mathrm{tr}}_{m,t+1} = 1$) and travels along the action-related road path $l^{\mathrm{c}}_{m,t+1} \in \mathcal{R}^{\mathrm{EVCS}}$ at $t+1$, which is updated by $l^{\mathrm{c}}_{m,t+1} = r(l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t})$. Additionally, the new arrival time is updated by $k^{\mathrm{a}}_{m,t+1} = t + \gamma_{l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t}} + 1$, where $\gamma_{l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t}}$ represent the traveling time between $l^{\mathrm{c}}_{m,t}$ and $l^{\mathrm{d}}_{m,t}$. In Case 2-2, the updates of road path $r(l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t})$ and traveling time $\gamma_{l^{\mathrm{c}}_{m,t}, l^{\mathrm{d}}_{m,t}}$ between $l^{\mathrm{c}}_{m,t}$ and $l^{\mathrm{d}}_{m,t}$ are performed using Dijkstra's algorithm. In Case 3, the MESS travels on the road in consecutive times ($b^{\mathrm{tr}}_{m,t} = b^{\mathrm{tr}}_{m,t+1} = 1$). Therefore, the arrival time and road path remain unchanged ($k^{\mathrm{a}}_{m,t+1} = k^{\mathrm{a}}_{m,t}$, $l^{\mathrm{c}}_{m,t+1} = l^{\mathrm{c}}_{m,t}$). Fig. 2 depicts the road routing of the three aforementioned cases based on Algorithm 1.

E. A SAFETY MODULE FOR MITIGATING SOC VIOLATION OF MESS

The SAC-based DRL agent performs a wide exploration during its training procedure to calculate its optimal action ($P^{\mathrm{agent}}_{b,m,t}$ in (31)) for the real charging and discharging power of MESSs. However, during the training procedure, the DRL agent may conduct unsafe exploration that yields overcharging (i.e., violation of the maximum SOC limit) and undercharging (i.e., violation of the minimum SOC limit) of a MESS. This degrades the performance of the MESS and reduces its lifetime. To resolve this issue, we propose a safety module for the DRL agent that enables the MESS to significantly reduce SOC violations through safe-exploration-based real charging and discharging power scheduling during the training procedure.

After the DRL agent selects its action ($P^{\mathrm{agent}}_{b,m,t}$) for MESS m at bus b and time t with the scheduling time unit Δt, the SOC of the MESS associated with the action is updated as follows.

$$SOC^{\mathrm{agent}}_{m,t} = SOC^{\mathrm{agent}}_{m,t-1} + \left(\frac{\eta^{\mathrm{ch}}_m P^{\mathrm{agent}}_{b,m,t} - \eta^{\mathrm{tr}}_m b^{\mathrm{tr}}_{m,t}}{E^{\mathrm{MESS,max}}_m} \right) \Delta t. \quad (40)$$

Subsequently, using the updated $SOC^{\mathrm{agent}}_{m,t}$ in (40), the SOC violation ($\Delta SOC^{\mathrm{v}}_{m,t}$) is calculated as follows.

$$\begin{aligned} \Delta SOC^{\mathrm{v}}_{m,t} &= (1 - b^{\mathrm{tr}}_{b,m,t}) \\ &\times \left\{ [\mathrm{sgn}(SOC^{\mathrm{min}}_m - SOC^{\mathrm{agent}}_{m,t})]^+ (SOC^{\mathrm{min}}_m - SOC^{\mathrm{agent}}_{m,t}) \right. \\ &\left. - [\mathrm{sgn}(SOC^{\mathrm{agent}}_{m,t} - SOC^{\mathrm{max}}_m)]^+ (SOC^{\mathrm{agent}}_{m,t} - SOC^{\mathrm{max}}_m) \right\} \end{aligned} \quad (41)$$

where $[x]^+$ and $\mathrm{sgn}(x)$ are denoted as $\max(x, 0)$ and $\frac{|x|}{x}$, respectively. Based on the SOC violation calculated using (41), the MESS real charging or discharging power ($P_{b,m,t}$) that eliminates its SOC violation is determined using the following equation.

$$P_{b,m,t} = P^{\mathrm{agent}}_{b,m,t} + \frac{E^{\mathrm{MESS,max}}_m \Delta SOC^{\mathrm{v}}_{m,t}}{\eta^{\mathrm{ch}}_m \Delta t}. \quad (42)$$

Finally, the value of the SOC between SOC^{min}_m and SOC^{max}_m is obtained using the following expression.

$$SOC_{m,t} = SOC^{\mathrm{agent}}_{m,t} + \frac{\eta^{\mathrm{ch}}_m \Delta P_{b,m,t} \Delta t}{E^{\mathrm{MESS,max}}_m} \quad (43)$$

where $\Delta P_{b,m,t}$ ensures that the real charging or discharging power of the MESS belongs to its allowable range $[-P^{\mathrm{max}}_{b,m}, P^{\mathrm{max}}_{b,m}]$, which is defined as follows.

$$\begin{aligned} \Delta P_{b,m,t} = P_{b,m,t} - P^{\mathrm{agent}}_{b,m,t} - [P_{b,m,t} - P^{\mathrm{max}}_{b,m}]^+ \quad (44) \\ + [-P^{\mathrm{max}}_{b,m} - P_{b,m,t}]^+. \quad (45) \end{aligned}$$

F. A SAFETY MODULE FOR MITIGATING NODAL VOLTAGE VIOLATION

The SAC-based DRL agent in charge of VVC using the MESS and PV system may lead to nodal voltage violations in power distribution systems through unsafe exploration with inadequate reactive power dispatch of the MESS and PV system during the training procedure. To mitigate such voltage violations, we propose the following iteration-based safety equations that update the reactive power absorption or injection from or to the grid for the MESS and PV system, respectively.

$$\begin{aligned} Q_{b,m,t}(k+1) &= Q_{b,m,t}(k) \\ &+ \rho^{\mathrm{MESS}}_{b,m,t} \left(Q^{\mathrm{agent}}_{b,m,t} - Q_{b,m,t}(k) \right) (1 - b^{\mathrm{tr}}_{m,t}) \quad (46) \end{aligned}$$

$$Q_{b,t}(k+1) = Q_{b,t}(k) + \rho^{\mathrm{PV}}_{b,t} \left(Q^{\mathrm{agent}}_{b,t} - Q_{b,t}(k) \right). \quad (47)$$

In (46) and (47), $\rho^{\mathrm{MESS}}_{b,m,t}$ and $\rho^{\mathrm{PV}}_{b,t}$ are adaptive parameters for the safe operation of the MESS and PV system in view of maintaining a normal voltage profile, respectively. $Q^{\mathrm{agent}}_{b,m,t}$ and $Q^{\mathrm{agent}}_{b,t}$ are the reactive powers (actions (32) and (33) of the DRL agent) of the MESS and PV system, respectively. Note that from (46), the safe reactive power adjustment of the MESS can be performed only when the MESS remains at the EVCS (i.e., $b^{\mathrm{tr}}_{m,t} = 0$). Using (46) and (47), the DRL agent iteratively updates the reactive power of the MESS and PV system based on their selected actions $Q^{\mathrm{agent}}_{b,m,t}$ and $Q^{\mathrm{agent}}_{b,t}$ to mitigate voltage violations. In particular, reactive powers of the MESS and PV system converge to $Q^{\mathrm{agent}}_{b,m,t}$ and $Q^{\mathrm{agent}}_{b,t}$, respectively, with increasing iterations k (i.e., $Q_{b,m,t}(k) = Q^{\mathrm{agent}}_{b,m,t}$ and $Q_{b,t}(k) = Q^{\mathrm{agent}}_{b,t}$) when the voltage magnitude is within the safe range $\left[V^{\mathrm{min}} + \epsilon, V^{\mathrm{max}} - \epsilon\right]$. In this range, ϵ indicates the safety margin, which is set with a small positive constant. The update of the reactive power is terminated when the maximum or minimum voltage magnitude remains in the safety margin region (i.e., $V^{\mathrm{max}} - \epsilon \leqslant V^{\mathrm{max}}(k) \leqslant V^{\mathrm{max}}$ and $V^{\mathrm{min}} \leq V^{\mathrm{min}}(k) \leq V^{\mathrm{min}} + \epsilon$). In summary, overvoltage and undervoltage are prevented by the aforementioned iterative update policy based on the safety-margin region when the

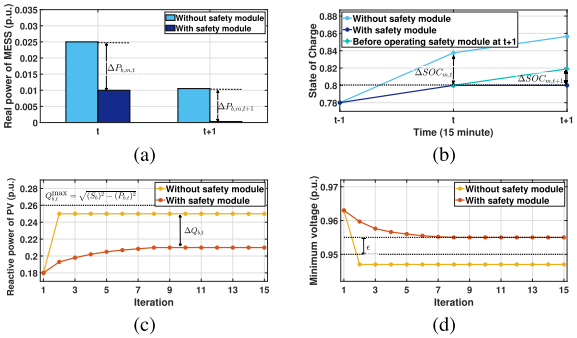


FIGURE 3. Illustrative examples of safety modules for the MESS and PV system with decreasing real and reactive power: (a) real power of MESS; (b) maximum SOC of MESS; (c) reactive power of PV system; and (d) minimum voltage magnitude.

maximum and minimum voltage magnitudes approach their limits, respectively.

During the iterations expressed in (46) and (47), $\rho^{\text{MESS}}_{b,m,t}$ and $\rho^{\text{PV}}_{b,t}$ are adaptively tuned according to the values of $V^{\max}(k)$ and $V^{\min}(k)$ at iteration k; they are expressed as follows.

$$\rho^{\text{MESS}}_{b,m,t} = \Gamma_m \Big([s^{\text{MESS}}_{b,m,t}]^+ [V^{\max} - \epsilon - V^{\max}(k)]^+ + [-s^{\text{MESS}}_{b,m,t}]^+ [V^{\min}(k) - V^{\min} - \epsilon]^+ \Big) \quad (48)$$

$$\rho^{\text{PV}}_{b,t} = \Gamma_b \Big([s^{\text{PV}}_{b,t}]^+ [V^{\max} - \epsilon - V^{\max}(k)]^+ + [-s^{\text{PV}}_{b,t}]^+ [V^{\min}(k) - V^{\min} - \epsilon]^+ \Big). \quad (49)$$

In (48) and (49), $s^{\text{MESS}}_{b,m,t}$ and $s^{\text{PV}}_{b,t}$ are defined as $\text{sgn}(Q^{\text{agent}}_{b,m,t}(k) - Q_{b,m,t}(0))$ and $\text{sgn}(Q^{\text{agent}}_{b,t}(k) - Q_{b,t}(0))$, respectively. Γ_m and Γ_b are positive constant parameters; $V^{\max(\min)}$ is the maximum (minimum) limit of the permissible voltage magnitude range; and $V^{\max(\min)}(k)$ is the maximum (minimum) voltage magnitude of the power distribution system at the k-th iteration.

Finally, the converged actual reactive power of the MESS ($Q^*_{b,m,t}$) through iterative equation (46) must be within its allowable range $[-Q^{\max}_{b,m,t}, Q^{\max}_{b,m,t}]$ where $Q^{\max}_{b,m,t} = \sqrt{(S_m)^2 - \left(P_{b,m,t}\right)^2}$; it is calculated as follows.

$$Q_{b,m,t} = Q^*_{b,m,t} - [Q^*_{b,m,t} - Q^{\max}_{b,m,t}]^+ + [-Q^{\max}_{b,m,t} - Q^*_{b,m,t}]^+. \quad (50)$$

Figs. 3 show conceptual diagrams that illustrate the operation of the proposed safety modules for mitigating SOC and voltage violations, corresponding to Figs. 3(a), (b) and Figs. 3(c) and (d), respectively. Figs. 3(a) and (b) compare the real power of MESS m and its SOC at bus b and times t and $t+1$ without and with a safety module, respectively. In the absence of the safety module, the real power charging action of the DRL agent is applied to the power distribution system at time t, which causes an SOC violation of its maximum limit (0.8) at time t. This violation is aggravated when the MESS further charges the real power at $t+1$. In contrast,

Algorithm 2 Online SAC Algorithm Integrated With MESS Road Routing and Safety Modules for the MESS and PV System

Initialize weights of neural networks ψ, ψ_{targ}, θ, ϕ and replay buffer $\mathcal{D}$;
repeat
  foreach *time step t* **do**
    Compute the action (30) of the agent: $\mathbf{a}_t = \tanh\left(\mu_\phi(\mathbf{s}_t) + \sigma_\phi(\mathbf{s}_t) \odot \xi_t\right)$
    Update $b^{\text{tr}}_{m,t+1}$, $k^{\text{a}}_{m,t+1}$, and $l^{\text{c}}_{m,t+1}$ using Algorithm 1;
    %Safe Exploration for MESS Charging/Discharging;
    Compute SOCs of MESSs using (40);
    Update real powers and SOCs of MESSs using (41)–(45);
    Compute the reactive powers of MESSs and PV systems using (32), (33);
    for *k=0* **to** *max* **do** *(%Safe Exploration for VVC)*
      Update reactive powers of MESSs and PV systems using (46), (47);
    end
    Obtain the new state s_{t+1} and the reward $r(s_t, a_t)$;
    Store a new tuple in the replay buffer $\mathcal{D}$. $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_t, a_t, r_t, s_{t+1}\}$
  end
  foreach *gradient step* **do**
    Perform a random sampling of a batch $\mathcal{B} \in \mathcal{D}$;
    Compute the gradient of three loss functions through (26)–(28);
    Update:
    $\psi \leftarrow \psi - \delta_V \nabla_{\boldsymbol{\psi}} J_V(\psi)$
    $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \delta_Q \nabla_{\boldsymbol{\theta}_i} J_Q(\boldsymbol{\theta}_i)$, with $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \delta_\pi \nabla_\phi J_\pi(\boldsymbol{\phi})$
    $\boldsymbol{\psi}_{\text{targ}} \leftarrow \tau\psi + (1-\tau)\boldsymbol{\psi}_{\text{targ}}$
  end
until *convergence*;

using the SOC safety module, the SOC violation is eliminated by reducing the real charging powers, $\Delta P_{b,m,t}$ and $\Delta P_{b,m,t+1}$, at times t and $t+1$, respectively, which are calculated using (45). Figs. 3(c) and (d) compare the reactive power of the PV system at bus b and the minimum voltage magnitude without and with a safety module, respectively. Without the safety module, a sudden absorption of the reactive power of the PV system leads to a voltage violation of its minimum limit (0.95 p.u.). However, the voltage safety module allows the PV system to absorb its reactive power while maintaining the voltage stability. In particular, according to the iterative equation (47) with the adaptive parameter (49), the reactive power absorption of the PV system significantly increases

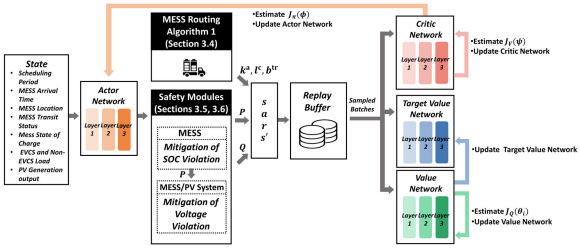


FIGURE 4. Architecture of the proposed SAC framework integrated with safety modules and MESS routing algorithm.

TABLE 1. Simulation parameters.

Category	Parameter	Value
MESS	$P_m^{\text{MESS,max}}$	140 kW
	$E_m^{\text{MESS,max}}$	500 kWh
	η_m^{tr}	0.001 kWh/Δt
	η_m^{ch}, η_m^{dch}	0.95
	$SOC_m^{\max(\min)}$	0.8(0.2)
	$SOC_{m,T}^{\text{tr}}$	0.5
	S_m	40 kVA
PV system	S_b^{PV}	350 kVA
Voltage magnitude	$V^{\max(\min)}$	1.05(0.95) p.u.
SAC	Batch size	256
	Replay buffer size	10^4
	Discount factor γ	0.2
	Soft update coefficient τ	0.5
	Temperature factor α	0.2
	Learning rate $\delta_V/\delta_Q/\delta_\pi$	$10^{-4}/10^{-4}/10^{-5}$
	No. hidden layers (all networks)	3
	Size of hidden layers	256
	Activation function	ReLU
	$\omega_1/\omega_2/\omega_3/\omega_4/\omega_5$	$10^3/2\times10^3/2\times 10^3/10^3/10^3$
Safety module	Γ_m/Γ_b	10/20
	ϵ	0.005

when the minimum voltage magnitude is much higher than its limit; thereupon, it slowly increases when the minimum voltage magnitude approaches its safety margin areas with ϵ. Subsequently, the reactive power and voltage magnitude become constant during iterations. The safety module for the MESS is implemented in the same way as the safety module for the PV system.

Finally, Algorithm 2 summarizes the procedure of the proposed SAC method integrated with the road routing algorithm for MESSs (Section III-D), and the safety modules of the MESSs and PV systems (Sections III-E and III-F). Fig. 4 shows the proposed SAC framework integrated with safety modules and the MESS routing algorithm in the CPT network.

IV. SIMULATION RESULTS

A. SIMULATION SETUP

The performance of the proposed SAC algorithm was analyzed in a coupled IEEE 33-bus power distribution [42] and

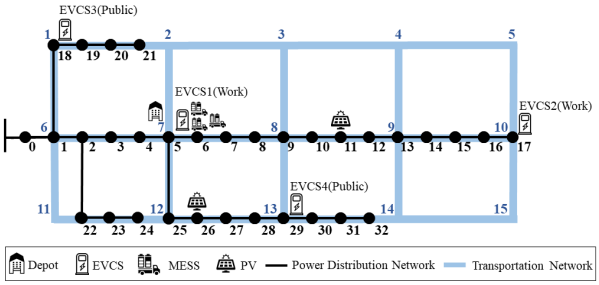


FIGURE 5. Coupled IEEE 33-bus power distribution and 15-node transportation systems.

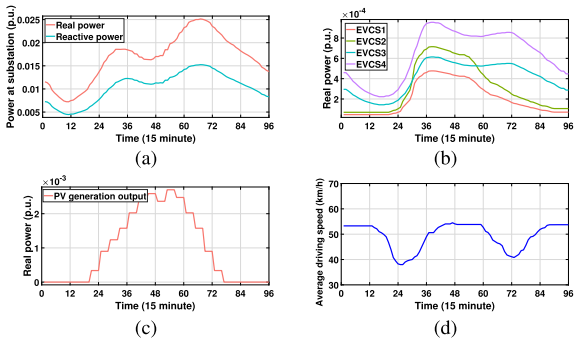


FIGURE 6. Simulation profile of power distribution and transportation systems during 24 h: (a) real and reactive power at substation; (b) real power consumptions of four EVCSs; (c) predicted output of PV real power generation; and (d) average driving speed of MESSs.

15-node transportation networks comprising three MESSs, two PV systems, and four EVCSs along with one depot, as shown in Fig. 5. The values of the simulation parameters are provided in Table 1. Figs. 6 show the total real and reactive power supply at the substation, aggregated real power EV loads of the four EVCSs, predicted output of PV real power generation, and average driving speed of the MESSs during the entire scheduling period. The proposed algorithm was trained and executed for 24 h ($T = 96$) with a 15-min scheduling resolution ($\Delta t = 15$ min). The proposed SAC and its corresponding MILP methods were simulated on an AMD Ryzen 7 with a 3700X Eight-Core Processor at 3.6 GHz and 32 GB of RAM using Python with the machine-learning package Pytorch and MATLAB R2020a with the IBM ILOG CPLEX Optimization Studio 12.8 solver, respectively.

B. TRAINING RESULTS

Figs. 7 compare two training curves of the total reward and negative real power loss between the SAC [41] (without safe exploration), CSAC [34] (with the SOC and voltage constraints), and proposed SAC methods (with the proposed safety modules) in the CPT network. Given that the SAC method has no safety module, the second and third reactive power mismatch terms in (37) and (36) for the MESSs and PV systems in the reward function of the proposed SAC method are replaced by the voltage mismatch term $\Delta V_{n,t}$ in the SAC method, where $\Delta V_{n,t}$ is the deviation of the voltage magnitude from its admissible range $[V^{\min}, V^{\max}]$. Note from

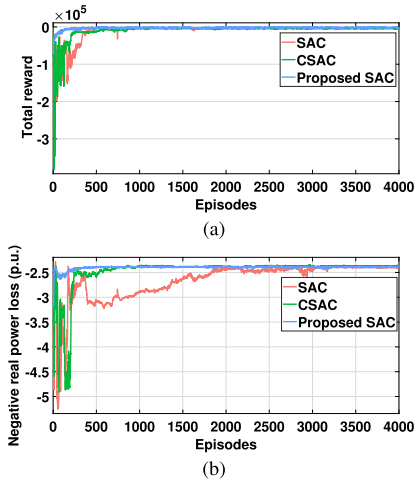


FIGURE 7. Comparison of training curves between the SAC, CSAC, and proposed SAC methods for: (a) total reward and (b) negative real power loss.

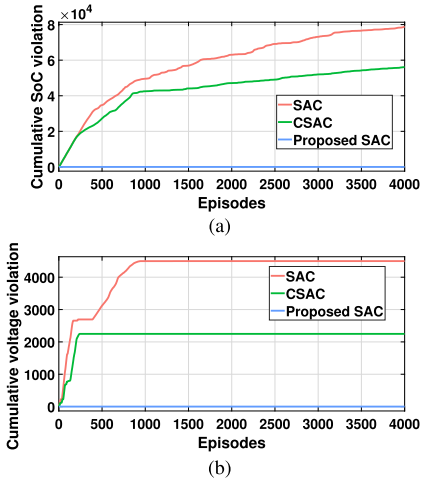


FIGURE 8. Comparison of cumulative violations between the SAC, CSAC, and proposed SAC methods in the coupled IEEE 33-bus power distribution and 15-node transportation systems for: (a) SOC of MESS and (b) voltage.

Fig. 7(a) that the training curve of the total reward for the proposed method converges faster than those of the other SAC and CSAC methods. This phenomenon occurs because the former negative real power loss included in the total reward increases more quickly in the early training period owing to the proposed safety modules than the latter negative real power loss without and with the conventional SOC and voltage constraints, as shown in Fig. 7(b).

Figs. 8 compare the cumulative violations of SOCs for the three MESSs and the voltage magnitude at any bus for the SAC, CSAC, and proposed SAC methods, respectively. In these figures, the SOC and voltage violation indicate the total number of MESSs and buses that reached values of the SOC and voltage magnitude beyond their acceptable ranges, i.e. [0.19, 0.8] and [0.95, 1.05] p.u. respectively, during the entire training period. Note that the original minimum limit of the MESS SOC was 0.2. However, the SOCs of MESSs decreased slightly owing to their travels after safe

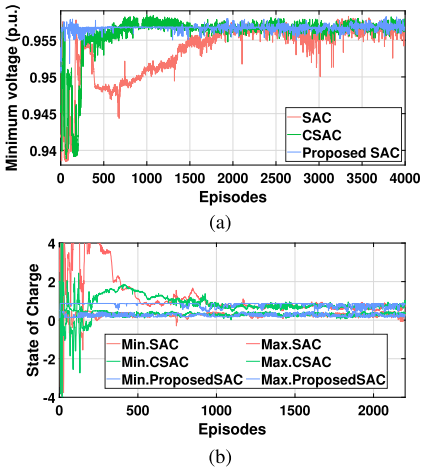


FIGURE 9. Performance comparison between the SAC, CSAC, and proposed SAC methods for: (a) minimum voltage magnitude and (b) minimum and maximum SOCs.

exploration. Therefore, the minimum limit of the MESS SOC was relaxed to 0.19 with a margin of 0.01. Note from Fig. 8(a) that no SOC violations occurred for the proposed SAC method, whereas the number of cumulative SOC violations for the SAC and CSAC methods increased rapidly in early stages of the training period and slowly after those stages. Another observation is that in comparison with the SAC method, the CSAC method leads to a greater reduction in SOC violations because the CSAC method considers the SOC constraints of MESSs during the training process; however, the CSAC method cannot eliminate SOC violations completely. Regarding the comparison of the three methods in terms of voltage violation, similar observations to those from Fig. 8(a) were also verified in Fig. 8(b): i) no voltage violations are identified in the proposed SAC method and ii) the CSAC method yields a greater reduction in voltage violations than the SAC method. The results in Figs. 8(a) and (b) demonstrate the effectiveness of the proposed safety modules, which eliminate SOC and voltage violations completely during the training process of the safety-integrated SAC DRL agent.

Table 2 summarizes the minimum/maximum values of the voltage magnitude and SOCs of the three MESSs at any bus along with their corresponding cumulative violations for the aforementioned three methods during the online training period. From this table, we point out the following three observations. First, the proposed SAC method resulted in no voltage violations with a minimum voltage magnitude of 0.9505 p.u., whereas the SAC and CSAC methods yielded significant voltage violations with minimum voltage magnitudes of 0.9385 and 0.9389 p.u., respectively. This observation was verified through Fig. 9(a), which shows that the proposed SAC method maintained a minimum voltage magnitude larger than 0.95 p.u. during the entire training period, whereas the SAC and CSAC methods generated a minimum voltage magnitude significantly less than 0.95 p.u., particularly in early stages of the training period. Second, no SOC

TABLE 2. Online training performance results in the coupled IEEE 33-bus power distribution and 15-node transportation systems.

Method	$V^{\min}$(p.u)	$V^{\max}$(p.u)	Voltage violations	$SOC^{\min}$	$SOC^{\max}$	SOC violations	Required SOC violations
SAC	0.9385	1	4489	-5.0053	5.9616	78969	3986
CSAC	0.9389	1	2248	-5.5531	6.1249	56073	2302
Proposed	**0.9505**	**1**	**0**	**0.1940**	**0.8**	**0**	**921**

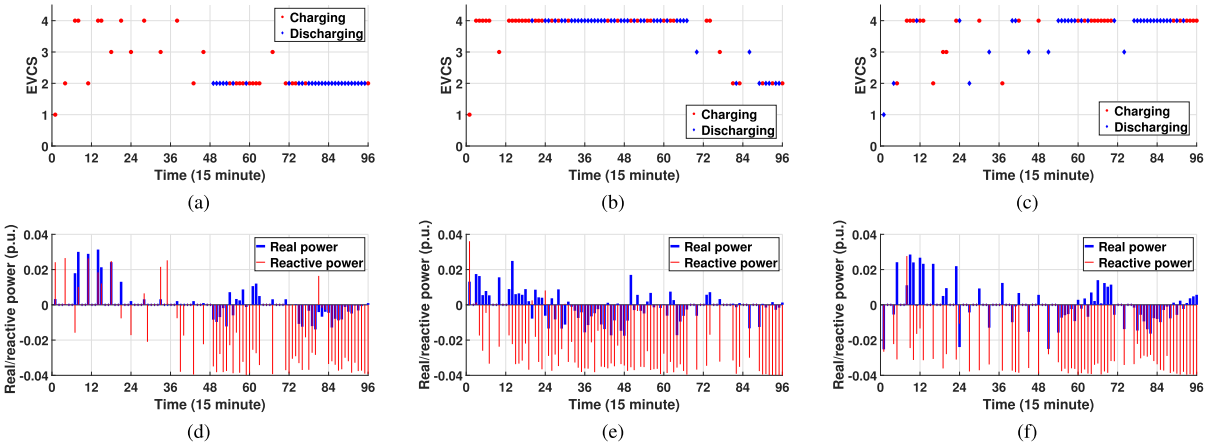


FIGURE 10. Routing and real/reactive power charging/discharging schedule of three MESSs using the proposed SAC method: (a) location of MESS 1; (b) location of MESS 2; (c) location of MESS 3; (d) charging/discharging of MESS 1; (e) charging/discharging of MESS 2; and (f) charging/discharging of MESS 3.

violations occurred in the proposed SAC method in which the minimum and maximum SOC values remained within the acceptable SOC range, i.e., [0.19, 0.8], corresponding to 0.194 and 0.8, respectively. Fig. 9(b) compares the minimum and maximum values of the SOCs for three MESSs between the SAC, CSAC, and proposed SAC methods. From this figure that, it can be verified that in comparison with the SAC and CSAC methods, which present large fluctuations in SOC during early stages of the training period, the proposed SAC method mitigates the SOC violations during such early stages and maintains the SOC within its permissible range. Third, the required SOC violations occurred in the training process of all three methods; however, the number of such violations was significantly reduced by the proposed SAC method. This is because the required SOC mismatch term belongs to the reward function of the proposed SAC method as a penalty and the range of the SOC is limited by the SOC safety module during the entire training period.

C. EXECUTION RESULTS

In this subsection, the execution performance of the proposed SAC method is analyzed and compared with those of the following four baseline methods: i) no VVC, ii) SAC, iii) CSAC, and iv) MILP-based optimization. The formulation of the MILP optimization method is based on the system model illustrated in Sections II-A–II-C along with the negative reward function, which is modified according to the formulation in [22]. Table 3 shows the average real power loss and number of voltage/SOC/required SOC violations of the proposed and four baseline methods during the execution process. Note from the second column of Table 3 that in contrast with no VVC, SAC, and CSAC methods, the proposed SAC method reduces the average real power loss by 28.1%, 4.5%, and 4.3%, respectively. In addition, the average real power loss of the proposed SAC method was slightly higher (2.8%) than that of the MILP optimization method. Note from the third, fourth, and fifth columns of Table 3 that no voltage/SOC/required SOC violations for the proposed SAC, MILP, and CSAC methods occurred in the execution stage, whereas the SAC method did yield some SOC and required SOC violations. This is because the SAC method includes neither an SOC constraint and nor its corresponding safety module for a safe exploration of the agent. This additionally leads to a slower convergence of the agent's training curve when the agent determines the complex routing paths of the MESSs in the transportation network.

TABLE 3. Execution performance results in the coupled IEEE 33-bus power distribution and 15-node transportation systems.

Method	Avg. real power loss (kW)	Voltage violations	SOC violations	Required SOC violations
NO VVC	3160	97	0	0
MILP	2207	0	0	0
SAC	2379	0	28	1
CSAC	2373	0	0	0
Proposed	**2272**	**0**	**0**	**0**

Figs. 10 compare the road routing and charging/discharging real/reactive power of the three MESSs during the entire scheduling period. The y-axis in Figs. 10(a)–(c) represent the indices of the four EVCSs where the MESSs remain and perform the charging and discharging processes. On the

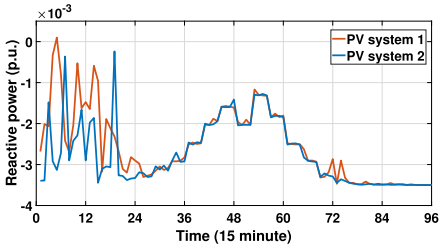


FIGURE 11. Reactive power schedule of PV systems using the proposed SAC method.

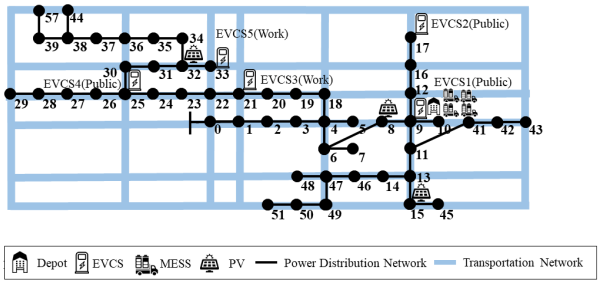


FIGURE 12. Coupled IEEE 57-bus power distribution and 42-node transportation systems.

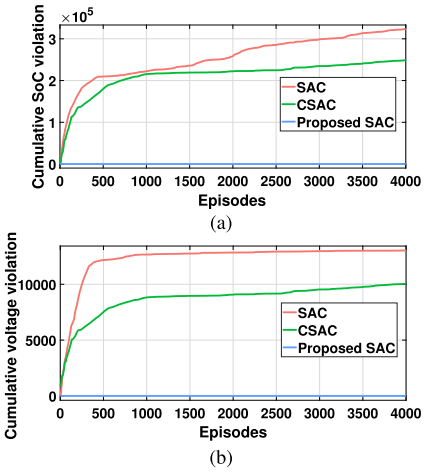


FIGURE 13. Comparison of cumulative violations between the SAC, CSAC, and proposed SAC methods in the coupled IEEE 57-bus power distribution and 42-node transportation systems for: (a) SOC of MESS and (b) voltage.

y-axis in Figs. 10(d)–(f), the positive and negative real/reactive power correspond to charging and discharging, respectively. Note from Figs. 10(a)–(c) that in general, the MESSs are dispatched to EVCSs 2 and 4, where the MESSs inject real and reactive power into the grid during their discharging process, as shown in Figs. 10(d)–(f). This is because the MESSs discharge their real and reactive power at EVCSs 2 and 4 further away from the substation to minimize the real power loss while maintaining a normal voltage profile. In addition, note from Figs. 10(d) and (f) that MESSs 1 and 3 tend to inject reactive power into the grid during the scheduling time periods [48, 96], when the grid has high loading conditions, as shown in Fig. 6(a). This is because the reactive power injection of these MESSs prevents undervoltage violations resulting from high load consumption. Furthermore, the MESSs recharge the real power from the grid to travel to another EVCS and discharge it to the EVs as shown in Figs. 10(a)–(c).

Fig. 11 shows the reactive power schedules of two PV systems using the proposed SAC method. Note from this figure that the amount of injected reactive power of the PV systems decreased and increased during the scheduling periods [36, 72] and [72, 96], respectively. This is because the total net power consumption (i.e., the difference between the power at the substation in Fig. 6(a) and the PV generation output in Fig. 6(c)) was low and high in the former and latter time periods, thereby injecting less and more reactive power into the grid to improve the real power loss reduction and voltage stability, respectively.

D. SCALABILITY

The scalability of the proposed SAC algorithm was validated in a coupled IEEE 57-bus power distribution [43] and 42-node transportation networks comprising four MESSs, three PV systems, and five EVCSs along with one depot, as shown in Fig. 12. The real and reactive power at substation was scaled up to be consistent with the IEEE 57-bus power distribution system' load data. The real power consumption of additional EVCS5 was scaled down from that of EVCS 1 in the IEEE 33-bus power distribution system. The identical simulation parameters in Table 1 were used in the coupled IEEE 57-bus power distribution and 42-node transportation networks except the apparent power $S_b^{\text{PV}} = 1200$ kVA. All training performance observations made from Table 2 and Figs. 8(a) and (b) (IEEE 33-bus power distribution system) can also be made from Table 4 and Figs. 13(a) and (b) (IEEE 57-bus power distribution system): i) no voltage and SOC violations and ii) reduction of the required SOC violations compared to the SAC and CSAC methods. In addition, similar to the execution performance results in Table 3 (IEEE 33-bus power distribution system), Table 5 (IEEE 57-bus power distribution system) shows that the proposed SAC method removes the voltage and SOC violations completely and yields slightly greater average real power loss than the benchmarked MILP optimization method during the execution process.

The novelty and valuable observations of the proposed SAC method are summarized as follows.

- To the best of authors knowledge, the proposed approach is the first SAC-based DRL framework (Algorithm 2) that jointly schedules the operation of MESSs and PV systems for VVC considering the road routing of MESSs (Algorithm 1) and safe exploration (Sections III-E and III-F) of the DRL agent in the CPT system.
- In comparison with conventional SAC and CSAC methods, the proposed safety-integrated SAC approach eliminates the MESS SOC and voltage violations completely during the entire training period of the DRL agent (see Figs. 7 and Table 2).

TABLE 4. Online training performance results in the coupled IEEE 57-bus power distribution and 42-node transportation systems.

Method	$V^{\min}$(p.u)	$V^{\max}$(p.u)	Voltage violations	$SOC^{\min}$	$SOC^{\max}$	SOC violations	Required SOC violations
SAC	0.9768	1.1051	13648	-5.8105	5.5645	333069	4164
CSAC	0.9769	1.0968	10031	-4.7813	5.4067	269805	2899
Proposed	**0.9770**	**1.0405**	**0**	**0.1939**	**0.8665**	**0**	**85**

TABLE 5. Execution performance results in the coupled IEEE 57-bus power distribution and 42-node transportation systems.

Method	Avg. real power loss (kW)	Voltage violations	SOC violations	Required SOC violations
NO VVC	5836	520	0	0
MILP	5371	0	0	0
SAC	5546	92	84	2
CSAC	5541	61	45	2
Proposed	**5526**	**0**	**0**	**0**

- Compared to no VVC, SAC, and CSAC methods, the proposed approach further reduces the average real power loss by 28.1%, 4.5%, and 4.3% during the execution period of the DRL agent, respectively, while no SOC and voltage violations occurred. The proposed approach yielded slightly greater average real power loss (2.8%) than the benchmarked MILP optimization method (see Table 3).

V. CONCLUSION

In this study, an SAC-based DRL framework integrated with a road routing algorithm of MESSs and plug-and-play safety modules of both MESSs and PV systems was proposed to jointly conduct the scheduling of MESS and PV system operations for VVC in a CPT network. Compared to conventional model-based optimization methods that address uncertainties inadequately along with a large computation time, the proposed framework is a model-free DRL method that is computationally efficient and robust in uncertain environment of the CPT network. Furthermore, compared to conventional safety-constrained DRL algorithms, the proposed safety modules need no modification of DRL structure so that they can be easily incorporated into any DRL algorithm. The proposed framework enables a DRL agent to minimize the real power loss and satisfy EV loads at EVCSs in the active power distribution network by i) dispatching the MESSs via the transportation network to EVCSs where their real/reactive power charging and discharging processes are performed, and ii) regulating the reactive power of MESSs and PV systems, while ensuring no MESS' SOC and voltage violations. The proposed framework was simulated in: i) coupled IEEE 33-bus power distribution and 15-node transportation systems comprising two PV systems, three MESSs, and four EVCSs and ii) coupled IEEE 57-bus power distribution and 42-node transportation systems comprising three PV systems, four MESSs, and five EVCSs. The simulation results demonstrated the superior performance of the proposed DRL approach over conventional DRL approaches in terms of real power loss, SOC/voltage violation, and convergence speed of the training curve of the DRL agent.

In the future, we will extend the proposed centralized DRL framework to a decentralized DRL framework, including conventional voltage regulating devices (e.g., OLTCs and CBs) to reduce the computational complexity. A key part of this future work is to develop a multi-agent DRL framework in which heterogeneous DRL agents of conventional voltage regulating devices and smart inverters of MESSs and PV systems cooperate to maintain stable power distribution grid operations and support EV loads in CPT networks.

REFERENCES

[1] X. Chang, Y. Xu, and H. Sun, "Vertex scenario-based robust peer-to-peer transactive energy trading in distribution networks," *Int. J. Electr. Power Energy Syst.*, vol. 138, pp. 1–11, Jun. 2022.
[2] K. E. Antoniadou-Plytaria, I. N. Kouveliotis-Lysikatos, P. S. Georgilakis, and N. D. Hatziargyriou, "Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2999–3008, Nov. 2017.
[3] S. Lakshmi and S. Ganguly, "An on-line operational optimization approach for open unified power quality conditioner for energy loss minimization of distribution networks," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 4784–4795, Nov. 2019.
[4] T. Terlouw, T. AlSkaif, C. Bauer, and W. van Sark, "Multi-objective optimization of energy arbitrage in community energy storage systems using different battery technologies," *Appl. Energy*, vol. 239, pp. 356–372, Apr. 2019.
[5] Y. Wang, T. Zhao, C. Ju, Y. Xu, and P. Wang, "Two-level distributed Volt/VAR control using aggregated PV inverters in distribution networks," *IEEE Trans. Power Del.*, vol. 35, no. 4, pp. 1844–1855, Aug. 2020.
[6] L. Wang, F. Bai, R. Yan, and T. K. Saha, "Real-time coordinated voltage control of PV inverters and energy storage for weak networks with high PV penetration," *IEEE Trans. Smart Grid*, vol. 33, no. 3, pp. 3383–3395, May 2018.
[7] Y. Hu, W. Liu, and W. Wang, "A two-layer Volt-VAR control method in rural distribution networks considering utilization of photovoltaic power," *IEEE Access*, vol. 8, pp. 118417–118425, 2020.
[8] C. Zhang, Y. Xu, Z. Dong, and J. Ravishankar, "Three-stage robust inverter-based voltage/var control for distribution networks with high-level PV," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 782–793, Jan. 2019.
[9] H. Lee, J.-C. Kim, and S.-M. Cho, "Optimal volt–var curve setting of a smart inverter for improving its performance in a distribution system," *IEEE Access*, vol. 8, pp. 157931–157945, 2020.
[10] X. Jiang, J. Chen, Q. Wu, W. Zhang, Y. Zhang, and J. Liu, "Two-step optimal allocation of stationary and mobile energy storage systems in resilient distribution networks," *J. Mod. Power Syst. Clean Energy*, vol. 9, no. 4, pp. 788–799, 2021.
[11] S. Yao, P. Wang, X. Liu, H. Zhang, and T. Zhao, "Rolling optimization of mobile energy storage fleets for resilient service restoration," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1030–1043, Mar. 2020.
[12] X. Liu, C. B. Soh, T. Zhao, and P. Wang, "Stochastic scheduling of mobile energy storage in coupled distribution and transportation networks for conversion capacity enhancement," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 117–130, Jan. 2021.
[13] W. Wang, X. Xiong, Y. He, J. Hu, and H. Chen, "Scheduling of separable mobile energy storage systems with mobile generators and fuel tankers to boost distribution system resilience," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 443–457, Jan. 2022.

[14] Y. Wang, A. O. Rousis, and G. Strbac, "Resilience-driven optimal sizing and pre-positioning of mobile energy storage systems in decentralized networked microgrids," *Appl. Energy*, vol. 305, pp. 1–13, Jan. 2022.

[15] G. Pulazza, N. Zhang, C. Kang, and C. A. Nucci, "Transmission planning with battery-based energy storage transportation for power systems with high penetration of renewable energy," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 4928–4940, Nov. 2021.

[16] S.-N. Yang, H.-W. Wang, C.-H. Gan, and Y.-B. Lin, "Mobile charging station service in smart grid networks," in *Proc. IEEE 3rd Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2012, pp. 1–6.

[17] S. Jeon and D.-H. Choi, "Optimal energy management framework for truck-mounted mobile charging stations considering power distribution system operating conditions," *Sensors*, vol. 21, no. 8, pp. 1–24, 2021.

[18] N. Chen, M. Li, M. Wang, J. Ma, and X. Shen, "Compensation of charging station overload via on-road mobile energy storage scheduling," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[19] H. Chen, Z. Su, Y. Hui, and H. Hui, "Dynamic charging optimization for mobile charging stations in Internet of Things," *IEEE Access*, vol. 6, pp. 53509–53520, 2018.

[20] Y. Zhang, X. Liu, W. Wei, T. Peng, G. Hong, and C. Meng, "Mobile charging: A novel charging system for electric vehicles in urban areas," *Appl. Energy*, vol. 278, pp. 1–7, Nov. 2020.

[21] M. Nazari-Heris, A. Loni, S. Asadi, and B. Mohammadi-Ivatloo, "Toward social equity access and mobile charging stations for electric vehicles: A case study in Los Angeles," *Appl. Energy*, vol. 311, pp. 1–15, Apr. 2022.

[22] S. Jeon and D.-H. Choi, "Joint optimization of Volt/VAR control and mobile energy storage system scheduling in active power distribution networks under PV prediction uncertainty," *Appl. Energy*, vol. 310, pp. 1–12, Mar. 2022.

[23] S. Wang, S. Bi, and Y. A. Zhang, "Reinforcement learning for real-time pricing and scheduling control in EV charging stations," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 849–859, Feb. 2021.

[24] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[25] G. Guo and Y. Gong, "Energy management of intelligent solar parking lot with EV charging and FCEV refueling based on deep reinforcement learning," *Int. J. Electr. Power Energy Syst.*, vol. 140, pp. 1–17, Sep. 2022.

[26] M. Alqahtani and M. Hu, "Dynamic energy scheduling and routing of multiple electric vehicles using deep reinforcement learning," *Energy*, vol. 244, pp. 1–11, Apr. 2022.

[27] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.

[28] B. Svetozarevic, C. Baumann, S. Muntwiler, L. D. Natale, M. N. Zeilinger, and P. Heer, "Data-driven control of room temperature and bidirectional EV charging using deep reinforcement learning: Simulations and experiments," *Appl. Energy*, vol. 307, pp. 1–16, Feb. 2022.

[29] B. Lin, B. Ghaddar, and J. Nathwani, "Deep reinforcement learning for the electric vehicle routing problem with time windows," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11528–11538, Aug. 2022.

[30] S. Yao, J. Gu, H. Zhang, P. Wang, X. Liu, and T. Zhao, "Resilient load restoration in microgrids considering mobile energy storage fleets: A deep reinforcement learning approach," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2020, pp. 1–5.

[31] Y. Wang, D. Qiu, and G. Strbac, "Multi-agent deep reinforcement learning for resilience-driven routing and scheduling of mobile energy storage systems," *Appl. Energy*, vol. 310, pp. 1–18, Mar. 2022.

[32] X. Sun and J. Qiu, "Two-stage Volt/VAR control in active distribution networks with multi-agent deep reinforcement learning method," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 2903–2912, Jul. 2021.

[33] D. Cao, J. Zhao, W. Hu, N. Yu, F. Ding, Q. Huang, and Z. Chen, "Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 149–165, Jan. 2022.

[34] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020.

[35] Y. Gao and N. Yu, "Model-augmented safe reinforcement learning for Volt-VAR control in power distribution networks," *Appl. Energy*, vol. 313, May 2022, Art. no. 118762.

[36] 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 Trans. Smart Grid*, vol. 13, no. 4, pp. 3266–3277, Jul. 2022.

[37] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based Volt-VAR optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, Jan. 2021.

[38] X. Sun and J. Qiu, "A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6852–6863, Oct. 2021.

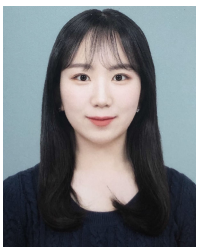
[39] H.-G. Yeh, D. F. Gayme, and S. H. Low, "Adaptive VAR control for distribution circuits with photovoltaic generators," *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1656–1663, Aug. 2012.

[40] S. Lei, J. Wang, C. Chen, and Y. Hou, "Mobile emergency generator pre-positioning and real-time allocation for resilient response to natural disasters," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2030–2041, May 2018.

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 1861–1870.

[42] W. H. Kersting, "Radial distribution test feeders," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 975–985, Aug. 1991.

[43] M. M. Ansari, C. Guo, M. S. Shaikh, N. Chopra, I. Haq, and L. Shen, "Planning for distribution system with grey wolf optimization method," *J. Electr. Eng. Technol.*, vol. 15, no. 4, pp. 1485–1499, Jul. 2020.



SOI JEON (Student Member, IEEE) received the B.S. degree in mathematics and the M.Sc. degree in electrical and electronics engineering from Chung-Ang University, Seoul, South Korea, in 2020 and 2022, respectively. She is currently an Associate Researcher with Hyundai Electric, Seongnam, South Korea. Her research interests include energy management systems, optimal scheduling of distributed resource, and deep reinforcement learning.

HOANG TIEN NGUYEN (Graduate Student Member, IEEE) received the B.Eng. degree (Highest Hons.) in electrical engineering from the Hanoi University of Science and Technology (HUST), Hanoi, Vietnam, in 2020. He is currently pursuing the M.S. degree in electrical and electronics engineering with Chung-Ang University, Seoul, South Korea. His research interests include safe learning, data-driven optimization, and the control of power systems.

DAE-HYUN CHOI (Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2002, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2008 and 2014, respectively. From 2002 to 2006, he was a Researcher with Korea Telecom (KT), Seoul, where he worked on designing and implementing home network systems. From 2014 to 2015, he was a Senior Researcher with LG Electronics, Seoul, where he developed home energy management systems. He is currently an Assistant Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul. His research interests include power system state estimation, electricity markets, cyber-physical security of smart grids, and the theory and application of cyber-physical energy systems. He received the Best Paper Award from 2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm), Tainan, Taiwan.