Towards Microgrid Resilience Enhancement via Mobile Power Sources and Repair Crews: A Multi-Agent Reinforcement Learning Approach

Yi Wang, *Member, IEEE,* Dawei Qiu, *Member, IEEE,* Fei Teng, *Senior Member, IEEE,* and Goran Strbac, *Member, IEEE*

***Abstract*—Mobile power sources (MPSs) have been gradually deployed in microgrids as critical resources to coordinate with repair crews (RCs) towards resilience enhancement owing to their flexibility and mobility in handling the complex coupled power-transport systems. However, previous work solves the coordinated dispatch problem of MPSs and RCs in a centralized manner with the assumption that the communication network is still fully functioning after the event. However, there is growing evidence that certain extreme events will damage or degrade communication infrastructure, which makes centralized decision making impractical. To fill this gap, this paper formulates the resilience-driven dispatch problem of MPSs and RCs in a decentralized framework. To solve this problem, a hierarchical multi-agent reinforcement learning method featuring a two-level framework is proposed, where the high-level action is used to switch decision-making between power and transport networks, and the low-level action constructed via a hybrid policy is used to compute continuous scheduling and discrete routing decisions in power and transport networks, respectively. The proposed method also uses an embedded function encapsulating system dynamics to enhance learning stability and scalability. Case studies based on IEEE 33-bus and 69-bus power networks are conducted to validate the effectiveness of the proposed method in load restoration.**

***Index Terms*—Mobile power sources, Repair crews, Microgrid resilience, Power-transport network, Hierarchical multi-agent reinforcement learning.**

NOMENCLATURE

A. Indices and Sets

$t \in T$	Index and set of time steps (hours)
$b \in B$	Index and set of electric buses
$l \in L$	Index and set of power lines
$d \in D$	Index and set of loads
$g \in DG$	Index and set of diesel generators (DGs)
$g \in PV$	Index and set of photovoltaic generation (PVs)
$g \in I_{eg}$	Index and set of mobile emergency generators (MEGs)
$k \in I_{es}$	Index and set of mobile energy storage systems (MESSs)
$j \in I_{rc}$	Index and set of repair crews (RCs)
$n \in N_{ms}$	Index and set of MESS stations (MSs)
$w \in N_{rc}$	Index and set of damaged components

B. Parameters

Δt	Time resolution (1 hour)
c_d^{ls}	Load shedding cost of load d (£/kWh)
$\overline{P}_g^{eg}$	Maximum active power of MEG g (kW)
$\underline{P}_g^{eg}$	Minimum active power of MEG g (kW)
$\overline{Q}_g^{eg}$	Maximum reactive power of MEG g (kVAR)
$\underline{Q}_g^{eg}$	Minimum reactive power of MEG g (kVAR)
$\overline{P}_k$	Power capacity of MESS k (kW)
S_k^{max}	Maximum SoC of MESS k (%)
η_k^{esc}	Charging efficiency of MESS k (%)
η_k^{esd}	Discharging efficiency of MESS k (%)
RT_w^{rc}	Time period required to repair component w
RS_j^{rc}	Resource capacity of RC j
rs_w^{rc}	Resources required to repair component w
$\overline{P}_g^{dg}$	Maximum active power of DG g (kW)
$\underline{P}_g^{dg}$	Minimum active power of DG g (kW)
$\overline{Q}_g^{dg}$	Maximum reactive power of DG g (kVAR)
$\underline{Q}_g^{dg}$	Minimum reactive power of DG g (kVAR)
$\tilde{P}_{g,t}^{pv}$	Active power capacity of PV g at time t (kW)
$\overline{S}_g^{pv}$	Apparent power capacity of PV g at time t (kVA)
$\overline{P}_{d,t}^{ed}$	Baseline of active load d at time t (kW)
$\overline{Q}_{d,t}^{ed}$	Baseline of reactive load d at time t (kVAR)
$\overline{V}$	Maximum permissible voltage (p.u.)
$\underline{V}$	Minimum permissible voltage (p.u.)
r_{bp}	Resistance of line (b,p) (p.u.)
x_{bp}	Reactance of line (b,p) (p.u.)
$\overline{S}_{bp}$	Capacity limit of line (b,p) (kVA)

C. Variables

$P_{g,n,t}^{eg}$	Active power generation of MEG g in MS n at time t (kW)
$Q_{g,n,t}^{eg}$	Reactive power generation of MEG g in MS n at time t (kVAR)
$P_{k,n,t}^{esc}$	Charging power of MESS k in MS n at time t (kW)
$P_{k,n,t}^{esd}$	Discharging power of MESS k in MS n at time t (kW)
$S_{k,t}^{es}$	State of Charge (SoC) of MESS k at time t
$u_{k,t}^{es}$	Binary indicating the scheduling status of MESS

This work was supported by two UK EPSRC projects: 'Integrated Development of Low-Carbon Energy Systems (IDLES): A Whole-System Paradigm for Creating a National Strategy' (project code: EP/R045518/ 1) and 'Technology Transformation to Support Flexible and Resilient Local Energy Systems' (project code: EP/T021780/1), and one Horizon Europe project: 'Reliability, Resilience and Defense technology for the griD' (Grant agreement ID: 101075714). *(Corresponding author: Dawei Qiu.)*

Yi Wang, Dawei Qiu, Fei Teng, and Goran Strbac are with the Department of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, U.K. (e-mail: {yi.wang18, d.qiu15, f.teng, g.strbac}@imperial.ac.uk).

k at time t (1 if charging, 0 if discharging)

$Re^{rc}_{j,w,t}$ Binary indicating if RC j is repairing component w at time t (1 if repairing, 0 otherwise)

$z^{rc}_{j,w,t}$ Binary indicating if the component w has been repaired by RC j at time t (1 if repaired, 0 otherwise)

$u_{i,n,t}$ Binary indicating the connection status of mobile unit i on node n at time t (1 if connected, 0 otherwise)

$P^{dg}_{g,t}$ Active power generation of DG g at time t (kW)

$Q^{dg}_{g,t}$ Reactive power generation of DG g at time t (kVAR)

$P^{ed}_{d,t}$ Restored active load d at time t (kW)

$Q^{ed}_{d,t}$ Restored reactive load d at time t (kVAR)

$P^{pv}_{g,t}$ Active power of PV g at time t (kW)

$Q^{pv}_{g,t}$ Reactive power of PV g at time t (kVAR)

$V_{b,t}$ Voltage magnitude at bus b at time t (p.u.)

$P_{bp,t}$ Active power flow of line (b, p) at time t (kW)

$Q_{bp,t}$ Reactive power flow of line (b, p) at time t (kVAR)

$P^{pv}_{g,t}$ Active power output of PV g at time t (kW)

$e_{b,t}$ Binary indicating the energized status of bus b at time t (1 if energized, 0 otherwise)

$y_{bp,t}$ Binary indicating the energized status of line (b, p) at time t (1 if energized, 0 otherwise)

$F_{bp,t}$ Virtual power flow through branch (b, p) at time t

$F^{s}_{a,t}$ Virtual output of black-start resource a at time t

I. Introduction

A. Background

Extreme events (e.g., hurricanes, storms, and earthquakes) featured by high impact and low probability can cause a catastrophic effect on power systems (e.g., severe power outages) [1]. Given the serious disruptions, the main goal of a resilient power system should be to maintain the supply continuity of essential loads (e.g., medical facilities and police stations), constituting a system load restoration problem [2]. As one emerging type of *distributed energy resources* (DERs), *mobile power sources* (MPSs) have been applied in power systems to coordinate with traditional mobile resource *repair crews* (RCs) for system load restoration due to their mobility and flexibility characteristics [3]. However, the deployment of these MPSs and RCs also creates new complications for the resilient operation of power systems, as they are transiting into a decentralized fashion characterized by quick responses as well as various system dynamics and uncertainties [4]. In this context, it is necessary to develop an efficient decentralized control scheme for the resilient coordination of MPSs and RCs in a complex power-transport network.

B. Literature Review

In the existing literature, the dispatches of MPSs and RCs are commonly solved as individual problems in a centralized manner. On one hand, in [5], [6], RC routing behaviors are formulated as a deterministic optimization problem to address the load restoration in the initial stage, while uncertainties associated with repair time and demand are realized in the later stages. On the other hand, mobile emergency generators (MEGs) have also demonstrated their advanced mobility in restoring essential loads [7], [8] via stochastic programming (SP) considering uncertain line outages. Additionally, mobile energy storage systems (MESSs) as the key technology of MPSs normally coordinate with MEGs for load restoration [9], while uncertain contingencies are handled via robust optimization (RO) [10] and SP [3]. However, prompt load restoration can be influenced by both the operability of a to-be-repaired branch and the availability of MPSs, while effective coordination between RCs and MPSs can significantly enhance the speed and quality of service restoration [11]. It is difficult to achieve the optimal load restoration solely based on RCs or MPSs, leading to the requirement for integrated methods that can coordinate available flexibility resources. To address this issue, there have been a few studies [11]–[13] investigating the coordinated effect between MPSs and RCs for system load restoration in coupled power-transport networks. However, the above studies are all deterministic given the difficulty of modeling the complex stochasticity of MPS and RC operations in a centralized manner. In [14], a two-stage stochastic pre-allocation model is proposed to coordinate MPSs and RCs, capturing the uncertainties associated with line outages and the influence of three different types of photovoltaics (PV) systems. The detailed transport network model capturing road congestion impacts is also missing in [11]–[14]. In [15], both MPSs and RCs are employed for the load restoration problem of an integrated power and hydrogen distribution network, where transport network and road congestion are captured.

Furthermore, the network reconfiguration of distribution networks is regarded as an efficient method and should be considered in the load restoration process [15]–[18]. In [17], [18], a time-efficient network reconfiguration model is proposed for the resilience enhancement of distribution networks. A virtual network with a set of radiality constraints is firstly introduced in [17], which can significantly improve the computational performance. In [18], a network reconfiguration strategy considering microgrids (MGs), substations, and unsupplied load islands is developed for black-start load restoration after extreme events. In [15], the multi-period network reconfiguration process is considered in an integrated power and hydrogen distribution network. To simulate more realistic load restoration process, a sequential black-start restoration model is proposed in [16], which can effectively capture the influence of black-start resources and dynamically concerns the sequence of recovered lines and buses.

Although the above model-based centralized optimization methods have been successfully applied to solve various MPS and RC resilience enhancement problems, the following challenges have to be addressed in the real-world environment. First, centralized optimization methods solely depend on a single control center, requiring prohibitive communication and computation resources. Furthermore, the centralized manner can be prone to single-point failure as all decisions are taken by this central controller [19], [20]. Second, since power systems are moving towards a decentralized fashion, it is typically intractable to acquire operation models and technical parameters of MPSs and RCs. Third, because the power and transport networks are highly dynamic and stochastic, it is hard to generalize an adaptive control scheme that accounts for various system uncertainties (e.g., renewables, demand, traffic volumes, etc.). As the system uncertainties can be characterized by many factors (e.g., weather conditions, energy usage behaviors, driving habits, etc.), it is difficult to even represent the uncertainty probability distributions exactly. Last but not least, even though the system models and uncertainties can be known, solving a scenario-based stochastic optimization problem for load restoration is normally time-consuming. Therefore, developing a control scheme for these decentralized MPSs and RCs becomes important and urgent [10].

Reinforcement learning (RL) [21], a data-driven and model-free method, may solve time-coupled decision-making problems by learning optimal policies through repeated interactions with the environment without any *prior* knowledge. As an online learning method, RL can effectively utilize the growing amount of data from the environment, capture various uncertainties, and adjust to different state conditions. Furthermore, well-trained policies can be directly deployed to the practical test process in milliseconds without solving an optimization problem. Recently, RL has been successfully applied to many resilient power system operation problems, such as load restoration [22], [23], voltage regulation [24], etc. However, the up-to-date literature investigating the application of RL to the dispatch problems of mobile resources towards MG load restoration is still limited. In [25], a single-agent reinforcement learning (SARL) method is applied to optimize the dispatch decisions of four MESSs for critical load restoration in MGs. However, applying SARL to a multi-agent setup may raise the scalability issue since the action space increases significantly with agent size, thereby costing computational time with one central agent [26]. To address this issue, the authors in [27] propose a multi-agent reinforcement learning (MARL) method that models each MESS as an individual agent for MG load restoration via a multi-agent formulation. However, there are still several difficulties that can be identified and are worthy of further efforts. First, the coordinated effect of MPSs and RCS is not investigated in the above paper. It is noted that effectively coordinating these decentralized MPSs and RCs is capable of achieving better MG load restoration performance. Second, at each time step, the routing and scheduling actions are computed simultaneously. However, MPSs and RCs cannot simultaneously make routing decisions in the transport network and scheduling decisions in the power network because the decision-making processes in the two networks are mutually exclusive. Therefore, developing an effective MARL-based method to separately compute transport routing action and power scheduling action is important for MPSs and RCs to enhance MG resilience.

C. Contributions

Based on the review of previous work on both model-based [3], [5]–[8], [10]–[13] and model-free [22]–[25], [27] methods, a significant research gap has been identified, which drives the motivation behind this paper: no previous work has developed a coordinated control scheme of MPSs and RCs, operating in a decentralized manner for MG load restoration, and employing a model-free decision-making framework at the same time. To fill this knowledge gap, following contributions are achieved:

1) Develop a novel decentralized framework for the coordinated decision-making problem of MPSs and RCs towards MG resilience enhancement with the objective of maximizing restored loads. This framework overcomes certain limitations of previous work: i) in contrast to [11]–[13] neglecting the tranport sectors of mobile sources, the detailed transport network model capturing road congestion impacts is accounted for; and ii) in contrast to [3], [5]–[8], [10] formulating the dispatch problem in a centralized manner, the decentralized co-dispatching behaviors of MPSs and RCs are captured that do not depend on the central commands, therefore avoiding the single-point failure and constructing the privacy perseverance.
2) Formulate the coordinated decision-making problem of MPSs and RCs as a *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP), wherein MPSs and RCs are regarded as the agents who can operate in a decentralized manner without following the central commands. In this context, the mobility and flexibility of MPSs and RCs can be fully explored inside the power-transport network towards system load restoration, in the meantime without knowing system models and uncertainty parameters.
3) Propose a novel hierarchical and hybrid MARL method to solve the Dec-POMDP. First, it learns a high-level (HL) policy that can direct MPSs and RCs in choosing between transport routing and power scheduling. Second, it learns a low-level (LL) hybrid policy that can capture MPSs' discrete routing and continuous scheduling actions. Additionally, RCs learn their routing and repairing selections in the HL via the same hierarchical structure, while learning their discrete routing and repairing actions in the LL via a conventional categorical policy. In this setting, the MPS and RC agents can both benefit from the hierarchical structure to learn the effective routing and scheduling/repairing actions separately rather than learning these two kinds of actions simultaneously, of which one becomes invalid. To further improve the scalability and stability of MARL policy, an abstracted embedded function capturing system dynamics is introduced during the training process.
4) Validate the superior performance of the proposed MARL method over the state-of-the-art model-based and model-free methods in MG load restoration. A generalized dispatch policy can be formulated and adapted to different sizes of MPSs and RCs and power-transport networks as well as system uncertainties.

The rest of the paper is organized as follows. Section II presents the general formulations of the utilized MPSs and RCs as well as the operational models of both transport and power networks. Sections III and IV introduce the Dec-POMDP formulation and the proposed H2MAPPO method, respectively. In Section V, case studies are carried out and discussed on two experimental environments. Section VI draws

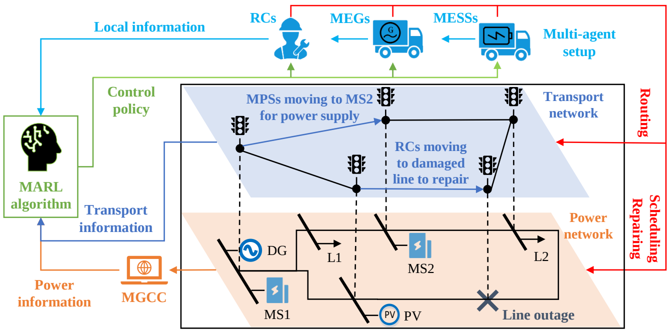


Fig. 1. The scheme of coordinated dispatch problem of MPSs and RCs in a power-transport network.

the conclusions and future work of this paper.

II. General Formulations of MPSs and RCs in Transport and Power Networks

A. Problem Descriptions

We focus on the resilience-driven coordinated dispatch problem of MPSs and RCs within a power-transport network including both routing and scheduling/repairing behaviors, as illustrated in Fig. 1. In general, electric components (e.g., buses and lines) in the power network are located on different transport nodes, while MPSs and RCs can move upon the transport network and choose to connect with their corresponding candidate nodes [7]. Specifically, we consider MEGs and MESSs as two types of MPSs that can choose to connect with the candidate nodes, e.g., MESS stations (MSs) [10]. Following [10], [14], this paper assumes that both MESSs and MEGs have black-start capability during the load restoration process. The role difference between MESSs and MEGs is that MESSs can charge power at one location with sufficient power supply and then discharge power at another location suffering load shedding, while MEGs can only provide power supply for the power network. In other words, MESSs play a role similar to demand-side response, whereas MEGs play a role similar to traditional generators but with mobility features. Regarding RCs, the candidate nodes are the initial depots and the locations of line outages. Inside the power network, static DERs, such as photovoltaics (PVs) and diesel generators (DGs), are installed suitably. In terms of the demand side, the power system captures both essential and non-essential loads to highlight the primary objective of load restoration [28].

Unlike the centralized framework [3], [5]–[8], [10]–[13], [25], this paper assumes that MPSs and RCs are operating in a decentralized manner, in anticipation of a future trend towards resilient distribution networks [10], [19], [20], [29]. In other words, MPSs and RCs can determine their individual dispatch behaviors without the central commands. In more detail, each resource can only acquire the local information of the coupled networks (e.g., PV generation, nodal load, line outage, and traffic volume) and its own status (e.g., transport location, battery state-of-charge (SoC), and repair capacity). Then, each resource determines its routing and scheduling/repairing decisions via an automatic control scheme. To solve the above decentralized coordination problem of MPSs and RCs, we first formulate this multi-agent setup as a Dec-POMDP, wherein MPSs and RCs are regarded as the agents and the coupled power-transport network is the environment. Then, a novel MARL method is proposed to drive MPS and RC agents to make optimal routing and scheduling/repairing actions in the transport and power networks respectively, aiming to maximize system resilience.

When the locations and repairs or power schedules of all mobile resources are settled, the *microgrid central controller* (MGCC) regulates each controlled DER and smart switches optimally in the power network for weighted load restoration maximization towards resilience enhancement. It is worth noting that the dispatch decisions of MPSs and RCs and the smart switch operations for network reconfiguration are mutually influenced. On one hand, the different locations and dispatch behaviors of MPSs and RCs can lead to different network reconfiguration results; on the other hand, MPS and RC agents repeatedly interact with the power network environment during the RL training process, gradually learn the key features of the environment (e.g., DER schedules and smart switch locations), and then adjust their dispatch behaviors towards better load restoration performance. In this section, we present the general formulations of the routing behaviors in the transport network and the scheduling/repairing behaviors in the power network.

B. Routing in Transport Network

The transport network is modeled as a weighted graph $G = \{N, R\}$, where $n \in N$ is the node set and $r \in R$ denotes the road set with the commuting distance L_r [25]. The general framework of MPSs and RCs' routing behaviors is illustrated in Fig. 2. Specifically, to commute in the transport network, they need to consider the influence of uncertain travel time and rely on an effective routing model, which are described in the following subsections.

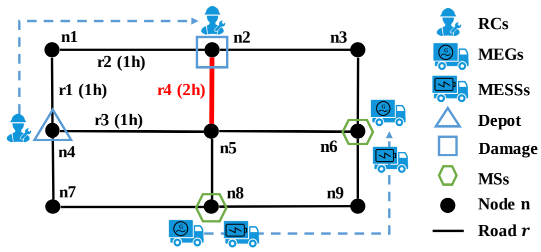


Fig. 2. Routing behaviors of MPSs and RCs in a transport network.

1) Uncertain Travel Time: The travel time $T_{r,t}^{rd}$ of road r is influenced by the real-time traffic volume [30], which can be estimated as

$$T_{r,t}^{rd} = \tilde{T}_r^{rd}[1 + \alpha^{rd}(\frac{V_{r,t}^{rd}}{C_r})^{\beta^{rd}}], \forall r \in R, \forall t \in T, \qquad (1)$$

where $\tilde{T}_r^{rd}$ is the free driving time of road r that mainly depends on the road distance, while C_r, α^{rd}, β^{rd}, and $V_{r,t}^{rd}$ correspond to the capacity, the retardation coefficients and the traffic volume of road r at time step t, respectively. Equation (1) can characterize the relationship between uncertain traffic volume and travel time, as well as reflect road impedance based on traffic flow itself. Note that heavy traffic volumes can cause serious road congestion and long travel time. It is crucial to account for the impact of road congestion on realistic routing behaviors; nevertheless, most existing literature on MPS and RC routing problems (e.g., [11]–[13]) ignores this factor. It is worth noting that the travel time between candidate nodes is directly associated with the dispatch behaviors of MPSs and RCs, while serious road congestion can significantly increase the required travel time. As such, ignoring road congestion impact can lead to inaccurate estimation of travel time and impractical dispatch behaviors of MPSs and RCs.

2) Routing Model: The routing behaviors of mobile units $i \in I$ including both MPSs and RCs within the transport network G can be formulated in a same manner [11], which are restricted by

$$\sum_{n\in N} u_{i,n,t} \leq 1, \forall i \in I, \forall t \in T, \tag{2}$$

$$\sum_{\tau=t}^{\min(t+T^{rd}_{mn,t},T)} u_{i,m,\tau} \leq (1-u_{i,n,t}) \cdot \min(T^{rd}_{mn,t}, T-t), \quad \forall i \in I, \forall n \in N, \forall m \in N \backslash \{i\}, \forall t \in T. \tag{3}$$

where constraint (2) shows that a mobile unit i can only be connected with one transport node for each time step, the binary variable $u_{i,n,t} \in \{0,1\}$ indicates the the mobile unit i connecting to the node n ($u_{i,n,t}=1$) or not ($u_{i,n,t}=0$) at time step t. Constraint (3) ensures reasonable routing behaviors of the mobile unit i between different transport nodes, where $T^{rd}_{mn,t}$ refers to the time period required to route from node n to node m at time step t. Note that $T^{rd}_{mn,t}$ can be uncertain and influenced by the real-time traffic volume $V^{rd}_{mn,t}$.

3) Routing Behaviors: In order to clearly illustrate the routing behaviors simulated in this paper, we take RCs in Fig. 2 as an example, the depot location of RCs is n_4 and the location of damaged line is n_2. Inside the network topology G, there are two available routes (i.e., $r_1 \rightarrow r_2$ and $r_3 \rightarrow r_4$) for RCs to commute, while serious congestion happening on road r_4 (the red road in Fig. 2) leads to much longer travel time ($T^{rd}_{r_{3+4},t} = 3$ hours) of route $r_3 \rightarrow r_4$. In this context, RCs will choose $r_1 \rightarrow r_2$ as their commuting route rather than $r_3 \rightarrow r_4$ due to the less 1 hour travel time. Regarding MEGs and MESSs, their routing behaviors in the network topology G can be derived in the similar manner as RCs. As shown in Fig. 2, the examined MEGs and MESSs are traveling from the initial MS at n_8 to the MS at n_6. Finally, when MPSs and RCs arrive at their destinations, they can be connected to the power network via MSs and damaged components, represented by N_{ms} and N_{rc} as subsets of N [11].

C. Scheduling/Repairing in Power Network

In power network, MPSs make scheduling decisions (i.e., MEG g power output $P^{eg}_{g,n,t}$, MESS k charging power $P^{esc}_{k,n,t}$ and discharging power $P^{esd}_{k,n,t}$) in MS n, RC j makes repairing decision $Re^{rc}_{j,w,t}$ for damaged component w. Afterwards, a linearized alternating current optimal power flow (AC-OPF) algorithm with the objective of load restoration maximization towards MG resilience enhancement can be solved by the MGCC for each time step t. The operation models of MEG, MESS, and RC as well as the employed AC OPF algorithm can be found in the following subsections.

1) MEG Scheduling: MEGs are constrained by their active and reactive power output limits

$$\underline{P}^{eg}_{g} \leq P^{eg}_{g,n,t} \leq \overline{P}^{eg}_{g}, \forall g \in I_{eg}, \forall n \in N_{ms}, \forall t \in T, \tag{4}$$

$$\underline{Q}^{eg}_{g} \leq Q^{eg}_{g,n,t} \leq \overline{Q}^{eg}_{g}, \forall g \in I_{eg}, \forall n \in N_{ms}, \forall t \in T. \tag{5}$$

where $P^{eg}_{g,n,t}$ and $Q^{eg}_{g,n,t}$ correspond to active and reactive power output of MEG g in MS n at time step t, respectively. We assume that adequate fuel is available for MPS routing through portable or towable fuel tanks and optimally dispatched fuel trucks in case of long term blackouts, following [3], [7], [10], [11], [14]. Additionally, to extend the continuous operating time of MPSs and handle the potential fuel issues, there are several potential solutions, e.g., appropriately selecting candidate nodes to connect MPSs and fuel trucks, deploying underground fuel tanks at candidate nodes, and preallocating fuel in the network [7], [11]. As such, the detailed logistic process based on the location of the re-filling station and the optimal re-filling time is not considered in this paper.

2) MESS Scheduling: The power operation model of MESSs can be appropriately presented via

$$0 \leq P^{esc}_{k,n,t} \leq u^{es}_{k,t}\overline{P}^{es}_{k}, \forall k \in I_{es}, \forall n \in N_{ms}, \forall t \in T, \tag{6}$$

$$-\overline{P}^{es}_{k}(1-u^{es}_{k,t}) \leq P^{esd}_{k,n,t} \leq 0, \forall k \in I_{es}, \forall n \in N_{ms}, \forall t \in T, \tag{7}$$

$$\underline{S}^{es}_{k} \leq S^{es}_{k,t} \leq \overline{S}^{es}_{k}, \forall k \in I_{es}, \forall t \in T, \tag{8}$$

$$S^{es}_{k,t+1} = S^{es}_{k,t} + \frac{P^{esc}_{k,n,t}\eta^{esc}_{k} + P^{esd}_{k,n,t}/\eta^{esd}_{k}}{\overline{E}^{es}_{k}}, \quad \forall k \in I_{es}, \forall n \in N_{ms}, \forall t \in T, \tag{9}$$

where constraints (6) and (7) restrict the maximum charging and discharging power of MESS k in MS n at time step t. The binary variable $u^{es}_{k,t} \in \{0,1\}$ introduced in constraints (6) and (7) indicates the charging ($u^{es}_{k,t}=1$) or discharging ($u^{es}_{k,t}=0$) behavior of MESS k at time step t. It is noted that these two behaviors cannot happen simultaneously. Constraint (8) limits the minimum and maximum battery SoC level of MESS k, while its dynamic transition between two consecutive time steps is presented in (9), given the charging/discharging power $P^{esc}_{k,t}, P^{esd}_{k,t}$ and efficiencies $\eta^{esc}_{k}, \eta^{esd}_{k}$.

3) RC Repairing: The repair plan of RC j is formulated as

$$z^{rc}_{j,w,t} \leq \frac{\sum_{\tau=1}^{t} Re^{rc}_{j,w,\tau}}{RT^{rc}_{w}}, \forall j \in I_{rc}, \forall w \in N_{rc}, \forall t \in T, \tag{10}$$

$$z^{rc}_{j,w,t} \leq z^{rc}_{j,w,t+1}, \forall j \in I_{rc}, \forall w \in N_{rc}, \forall t \leq T-1, \tag{11}$$

$$\sum_{w\in N_{rc}} rs^{rc}_{w} \cdot z^{rc}_{j,w,T} \leq RS^{rc}_{j}, \forall j \in I_{rc}. \tag{12}$$

where the binary variable $z^{rc}_{j,w,t}=1$ if the damaged component w is repaired by RC j at time step t, and $z^{rc}_{j,w,t}=0$ otherwise. Binary $Re^{rc}_{j,w,t}$ represents if RC j is repairing component w at time step t (1 if repairing, 0 otherwise), while RT^{rc}_{w} corresponds to the time period required to repair component w [11]. Constraint (12) ensures that the resource capacity RS^{rc}_{j}

of RC j is sufficient for its repair tasks, where rs_w^{rc} refers to the number of resources required to repair damaged component w.

4) Power Network: The power network operation is fully modeled by a linearized OPF algorithm for MG resilience enhancement. Specifically, once MEG g, MESS k and RC j are moving to their individual destinations, where MEG g and MESS k are making their power schedules $P_{g,t}^{eg}$ and $P_{k,t}^{es}$, while RC j is making its repairing decision $Re_{j,w,t}^{rc}$ for damaged component w. The following linearized OPF algorithm can be then solved by the MGCC for each time step t.

$$\Big\{ \max_{\Xi^{mg}} \mathbb{E}\{ \sum_{d\in D} c_d^{ls} P_{d,t}^{ed} \}, \tag{13}$$

where

$$\begin{aligned}\Xi^{mg} = \{&P_{d,t}^{ed}, Q_{d,t}^{ed}, P_{g,t}^{dg}, Q_{g,t}^{dg}, P_{g,t}^{pv}, Q_{g,t}^{pv},\\ &P_{bp,t}, Q_{bp,t}, V_{b,t}^2, y_{bp,t}, e_{b,t}, F_{bp,t}\},\end{aligned} \tag{14}$$

subject to

$$\begin{aligned}&\sum_{g\in B_{dg}} P_{g,t}^{dg} + \sum_{g\in B_{eg}} P_{g,t}^{eg} + \sum_{g\in B_{pv}} P_{g,t}^{pv} = \sum_{d\in B_{ed}} P_{d,t}^{ed}\\ &+ \sum_{k\in B_{es}} P_{k,t}^{es} - \sum_{(p,b)\in L} P_{pb,t} + \sum_{(b,p)\in L} P_{bp,t},\ \forall b\in B,\end{aligned} \tag{15}$$

$$\begin{aligned}&\sum_{g\in B_{dg}} Q_{g,t}^{dg} + \sum_{g\in B_{eg}} Q_{g,t}^{eg} + \sum_{g\in B_{pv}} Q_{g,t}^{pv} = \sum_{d\in B_{ed}} Q_{d,t}^{ed}\\ &- \sum_{(p,b)\in L} Q_{pb,t} + \sum_{(b,p)\in L} Q_{bp,t},\ \forall b\in B,\end{aligned} \tag{16}$$

$$\underline{P}_g^{dg} \le P_{g,t}^{dg} \le \overline{P}_g^{dg},\ \forall g\in DG, \tag{17}$$

$$\underline{Q}_g^{dg} \le Q_{g,t}^{dg} \le \overline{Q}_g^{dg},\ \forall g\in DG, \tag{18}$$

$$0 \le P_{g,t}^{pv} \le \tilde{P}_{g,t}^{pv},\ \forall g\in PV^{bs}, \tag{19}$$

$$(P_{g,t}^{pv})^2 + (Q_{g,t}^{pv})^2 \le (\overline{S}_g^{pv})^2,\ \forall g\in PV^{bs}, \tag{20}$$

$$0 \le P_{g,t}^{pv} \le e_{b,t}\tilde{P}_{g,t}^{pv},\ \forall b\in B,\ \forall g\in B_{pv}\cap PV^{nbs}, \tag{21}$$

$$(P_{g,t}^{pv})^2 + (Q_{g,t}^{pv})^2 \le e_{b,t}(\overline{S}_g^{pv})^2, \forall b\in B, \forall g\in B_{pv}\cap PV^{nbs}, \tag{22}$$

$$P_{d,t}^{ed} \le e_{b,t}\overline{P}_{d,t}^{ed},\ \forall b\in B, \forall d\in B_{ed}, \tag{23}$$

$$Q_{d,t}^{ed} \le e_{b,t}\overline{Q}_{d,t}^{ed},\ \forall b\in B, \forall d\in B_{ed}, \tag{24}$$

$$e_{b,t}\underline{V}^2 \le V_{b,t}^2 \le e_{b,t}\overline{V}^2,\ \forall b\in B, \tag{25}$$

$$P_{bp,t}^2 + Q_{bp,t}^2 \le y_{bp,t}\cdot \overline{S}_{bp},\ \forall (b,p)\in L, \tag{26}$$

$$\begin{aligned}V_{b,t}^2 - V_{p,t}^2 \le\ & 2\cdot (r_{bp}P_{bp,t} + x_{bp}Q_{bp,t})\\ &+(1-y_{bp,t})\cdot M,\ \forall (b,p)\in L,\end{aligned} \tag{27}$$

$$\begin{aligned}V_{b,t}^2 - V_{p,t}^2 \ge\ & 2\cdot (r_{bp}P_{bp,t} + x_{bp}Q_{bp,t})\\ &+(y_{bp,t}-1)\cdot M,\ \forall (b,p)\in L,\end{aligned} \tag{28}$$

$$\sum_{(b,p)\in L} y_{bp,t} = |B| - \big(\sum_{b\in B}(1-e_{b,t}) + N_{bs}\big), \tag{29}$$

$$\sum_{a\in B_{bs}} F_{a,t}^s + \sum_{(p,b)\in L} F_{pb,t} - \sum_{(b,p)\in L} F_{bp,t} = e_{b,t},\ \forall b\in B, \tag{30}$$

$$-y_{bp,t}\overline{F}_{bp} \le F_{bp,t} \le y_{bp,t}\overline{F}_{bp},\ \forall (b,p)\in L, \tag{31}$$

$$y_{bp,t} \le z_{j,bp,t}^{rc},\ \forall (b,p)\in N_{rc},\ \Big\},\quad t\in T, \tag{32}$$

where the objective function (13) is to maximize the expectation of weighted sum of restored loads [11]–[13] capturing various uncertainties (e.g., demand and PV generation) and stochastic variables in set Ξ^{mg}. The OPF constraints formulated by the LinDistFlow model [31] include the active and reactive power balances (15)-(16) at bus b, where B_{eg}, B_{es}, B_{ed}, B_{dg}, and B_{pv} correspond to the sets of MEG g, MESS k, restored load d, DG g and PV g located at bus b, respectively. The active and reactive power outputs of DG g are constrained by (17) and (18) respectively, while the output limits of grid-forming and grid-following PVs are presented in (19)-(20) and (21)-(22) respectively [32]. The status of load d is restricted by constraint (23) and (24), where binary $e_{b,t}$ corresponds to the energized status of bus b (1 if energized, 0 otherwise). The voltage and power flow limits are shown in (25) and (26) respectively, while the linearized power flow constraints are expressed in (27)-(28). Binary $y_{bp,t}$ in (26)-(28) indicates the energized status of line (b,p) (1 if energized, 0 otherwise) determined by both reconfiguration switch and RC operations, while M refers to a large positive number used to relax constraints (27)-(28) for disconnected or damaged lines.

To coordinate with mobile source dispatch, the power network can be dynamically reconfigured through smart switch operations, while this process should respect the system radiality, subject to a set of virtual network constraints (29)-(32) [14], [16]. According to [15], [17], [18], two conditions should be satisfied for the network radiality: i) each island is connected; ii) the number of energized lines is equal to the number of buses minus the number of islands. Furthermore, an additional condition is introduced by [16] to ensure that de-energized or cut-off buses are not connected with each other due to the lack of black-start resources. Specifically, constraint (29) maintains the network radiality during the network reconfiguration process, which shows that the number of lines is equal to the number of buses minus the number of possible islands containing black-start resources and de-energized buses [16]. Constraint (30) refers to the nodal power balance of the virtual network, indicating that bus b is energized if the virtual load at the bus is served ($x_{b,t} = 1$). B_{bs} is the set of power sources with black-start capabilities (e.g., buses connected with DGs, grid-forming PVs, and MPSs in the real power network) located at bus b. Constraint (31) models the connections between the power network and the virtual network through the line connection status $y_{bp,t}$. Constraint (32) indicates that the damaged line $(b,p)\in N_{rc}$ could be energized in the power network after being repaired by RC j ($z_{j,bp,t}^{rc} = 1$). More detailed mathematical formulations of the power network model can be found in [11], [14], [16]–[18].

III. Reformulation as a Dec-POMDP

The conventional method lies in using optimization approach to solve above coordinated dispatch problem of MPSs and RCs [11]–[13]. However, it is hard to generalize an adaptive and fast control scheme that accounts for various system dynamics and uncertainties in the context of power and transport networks, as discussed in Section I. To this end, it is reasonable to formulate this problem as a *Decentralized Partially Observable Markov Decision Process* (Dec-POMDP) [33], considering that MPSs and RCs are operating in a decentralized manner and can only observe partial information of the power and transport networks.

A Dec-POMDP is a 7-tuple $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, \mathcal{R}, \mathcal{T}, \gamma \rangle$, including $|\mathcal{I}|$ agents, a collection of global states $s \in \mathcal{S}$, local observations $\{o_i \in \mathcal{O}_{1:I}\}$, action sets $\{a_i \in \mathcal{A}_{1:I}\}$, and reward functions $\{r_i \in \mathcal{R}\}$, as well as a state transition function $\mathcal{T}(s, o_{1:I}, a_{1:I}, \omega)$, where ω is the environment stochasticity representing the system uncertain parameters. The time interval Δt is 1 hour. For each agent i at time step t, an action $a_{i,t}$ is computed using the policy $\pi_i(a|o)$ conditioned on the current local observation $o_{i,t}$. Then, the environment transits to the next state given the transition function $\mathcal{T}$, while agent i is rewarded $r_{i,t}$ and updated a next local observation $o_{i,t+1}$. Following this process, each agent i receives a trajectory of local observations, actions, and rewards: $\tau_i = o_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, ..., r_{i,T}$ over $\mathcal{O}_i \times \mathcal{A}_i \times \mathcal{R} \rightarrow \mathbb{R}$. The objective of each agent i is maximizing its cumulative discounted reward $R_i = \sum_{t=0}^{T} \gamma^t r_{i,t}$, where $\gamma \in [0, 1)$ and T = 24 hours are the discount factor and daily horizon, respectively. The components of Dec-POMDP are as below:

A. Agent

The agents are defined as three groups of resources: 1) RC agent $i = j \in I_{rc} \subseteq I$; 2) MEG agent $i = g \in I_{eg} \subseteq I$; and 3) MESS agent $i = k \in I_{es} \subseteq I$.

B. Environment

The environment can be organized into two sectors: 1) routing process of RC, MEG and MESS agents in the transport network; 2) repairing/scheduling process of RC, MEG and MESS agents in the power network. The routing constraints of mobile sources (2) and (3) can be satisfied automatically by realistic transport network settings (e.g., transport nodes, road distance, and traffic volumes) in the RL environment.

C. Observation

Each agent i at time step t observes its local observation $o_{i,t}$ differing for distinct groups, defined as

$$o_{i,t} = \begin{cases} [N_{j,t}, V^{rd}_{j,t}, S^{ln}_{j,t}, RT^{rc}_{j,t}, RS^{rc}_{j,t}, S^{rc}_{j,t}] & \forall i = j \in I_{rc} \\ [N_{g,t}, V^{rd}_{g,t}, S^{ln}_{g,t}, P^{ed}_{g,t}, P^{pv}_{g,t}] & \forall i = g \in I_{eg} \\ [N_{k,t}, V^{rd}_{k,t}, S^{ln}_{k,t}, P^{ed}_{k,t}, P^{pv}_{k,t}, S^{es}_{k,t}] & \forall i = k \in I_{es} \end{cases}, \tag{33}$$

comprising two parts: 1) the transport node index $N_{i,t}$ the agent i is traveling on and the corresponding road traffic volume $V^{rd}_{i,t}$; 2) the power information of the line status (outage or not) $S^{ln}_{i,t}$, nodal load $P^{ed}_{i,t}$, PV generation $P^{pv}_{i,t}$, the battery SoC $S^{es}_{i,t}$ of MESS agent i, the time $RT^{rc}_{i,t}$ and resources $RS^{rc}_{i,t}$ required to repair the damaged line as well as the current resource status $S^{rc}_{i,t}$ of RC agent i.

This paper assumes that agents can get access to the above required local information through interactions with the environment. However, it is worth noting that there exists a risk that these agents can only get access to the incomplete knowledge (or even no information) of the distribution network due to damaged communication infrastructure or data privacy concerns [34]. In this context, to discover accurate nodal information only using incomplete knowledge, the following two methods may be useful: 1) instead of directly observing the detailed nodal knowledge of the coupled system, agents can choose to acquire aggregated information (e.g., aggregated load profiles) or estimated information (e.g., solar irradiance) [34]; 2) agents can be equipped with effective forecasting mechanisms (e.g., long short-term memory (LSTM) networks [35]), which takes the incomplete information as input and then output the required local information for the RL training process.

D. Action

Each agent i at time step t controls its action $a_{i,t}$ that varies for different agent groups, defined as

$$a_{i,t} = \begin{cases} [a^{l,rc}_{j,t}, a^{r,rc}_{j,t}] & \forall i = j \in I_{rc} \\ [a^{l,eg}_{g,t}, a^{p,eg}_{g,t}] & \forall i = g \in I_{eg} \\ [a^{l,es}_{k,t}, a^{p,es}_{k,t}] & \forall i = k \in I_{es} \end{cases}, \tag{34}$$

comprising two parts: 1) discrete routing action $a^{l}_{i,t} \in \{0, 1, ..., R^{rd}\}$ is selected from the set of potential routes upon the transport node, in which 0 denotes no routing behaviors and R^{rd} denotes the number of available commuting routes at current transport node $N_{i,t}$, as described in [36]–[38]; 2) discrete action $a^{r,rc}_{j,t} \in \{0, 1\}$ corresponds to the choice of RC j to repair the component ($a^{r,rc}_{j,t} = 1$) or not ($a^{r,rc}_{j,t} = 0$); 3) continuous actions $a^{p,eg}_{g,t} \in [0, 1]$ and $a^{p,es}_{k,t} \in [-1, 1]$ represent the magnitude of power generation (for MEG agent g) and power charging/discharging (for MESS agent k) as a percentage of their power capacity $[\underline{P}^{eg}_{g}, \overline{P}^{eg}_{g}]$ and $[-\overline{P}^{es}_{k}, \overline{P}^{es}_{k}]$, respectively.

E. State Transition

The state transition process from time step t to $t+1$ is governed by $s_{t+1} = \mathcal{T}(s_t, o_{1:I,t}, a_{1:I,t}, \omega_t)$, which is affected by a combination of the current state s_t of environment, local observations $o_{1:I,t}$ and actions $a_{1:I,t}$ of agents, and environment stochasticity ω_t. Regarding this problem, the environment stochasticity $\omega_t = [S^{ln}_{l,t}, P^{ed}_{d,t}, P^{pv}_{g,t}, V^{rd}_{r,t}]$ corresponds to the exogenous states that are independent of agent actions and have intrinsic variability. RL can overcome this variability by adopting a data-driven fashion that does not depend on precise probability distributions for various uncertainties but instead learns state features from the data set itself [21]. To better prove the effectiveness of RL on handling environment uncertainties, a test dataset (separate from training dataset) is normally used to evaluate the performance of the trained RL policy on generalization to different state conditions. Furthermore, once the RL policy is well trained, it can be directly deployed to the practical test process in milliseconds.

On the other hand, the transition of endogenous states $N_{i,t}, RT_{i,t}^{rc}, RS_{i,t}^{rc}, S_{i,t}^{rc}, S_{i,t}^{es}$ can be determined by the agents' action $a_{i,t}$. Specifically, $N_{i,t}$ is determined by $a_{i,t}^{l}$, corresponding to the routing decisions upon the transport network, as expressed in Section II-B. Furthermore, as an RC agent i, once moving to a damaged line, the repairing time $RT_{i,t}^{rc}$ and resources $RS_{i,t}^{rc}$ of this line can be observed. If RC agent i decides to repair this line ($a_{i,t}^{r,rc} = 1$), the remaining resources $S_{i,t}^{rc}$ at time step t can be updated after the line is repaired:

$$S_{i,t+1}^{rc} = S_{i,t}^{rc} - a_{i,t}^{r,rc} RS_{i,t}^{rc}, \forall i \in I_{rc}. \tag{35}$$

Finally, as a storage unit, $S_{i,t}^{es}$ of MESS agent i is managed by its continuous action $a_{i,t}^{p,es}$ through the mutually exclusive quantities $P_{i,t}^{esc}, P_{i,t}^{esd}$, which are limited by its minimum and maximum SoC level $\underline{S}_{i}^{es}, \overline{S}_{i}^{es}$, energy and power capacities $\overline{E}_{i}^{es}, \overline{P}_{i}^{es}$, and charging/discharging efficiency η_{i}^{es}, depicted as

$$P_{i,t}^{esc} = [\min(a_{i,t}^{p,es}\overline{P}_{i}^{es}, (\overline{S}_{i}^{es} - S_{i,t}^{es})\overline{E}_{i}^{es}/\eta_{i}^{es}]^{+}, \forall i \in I_{es}, \tag{36}$$

$$P_{i,t}^{esd} = [\max(a_{i,t}^{p,es}\overline{P}_{i}^{es}, (\underline{S}_{i}^{es} - S_{i,t}^{es})\overline{E}_{i}^{es}\eta_{i}^{es}]^{-}, \forall i \in I_{es}, \tag{37}$$

where operator $[\cdot]^{+/-} = \max/\min\{\cdot, 0\}$. Then, the state transition of $S_{i,t}^{es}$ from time step t to $t+1$ is written as

$$S_{i,t+1}^{es} = \begin{cases} S_{i,t}^{es} + \frac{P_{i,t}^{esc}\eta_{i}^{es} + P_{i,t}^{esd}/\eta_{i}^{es}}{\overline{E}_{i}^{es}} & \text{if } a_{i,t}^{l,es} = 0 \\ S_{i,t}^{es} & \text{otherwise} \end{cases}, \forall i \in I_{es}, \tag{38}$$

where $a_{i,t}^{l,es} = 0$ indicates that the MESS agent i connects to the grid at time step t.

F. Reward

After the dispatches of all MPSs and RCs have been obtained, load restoration quantity $P_{d,t}^{ed}$ of each load d at time step t can be optimized via the linearized AC-OPF algorithm described in Section II-C. It is noted that all the constraints inside the distribution network can be satisfied by solving the optimization (13)-(32) with sufficient flexibility supported by DERs, load shedding, and switch switch operations. The studied problem aims at maximizing the weighted sum of restored loads for resilience enhancement. Thus, the reward function of agents in Dec-POMDP is designed as the following resilience index

$$r_{i,t} = \lambda_t = \frac{\sum_{d \in D} c_{d}^{ls} P_{d,t}^{ed}}{\sum_{d \in D} c_{d}^{ls} \overline{P}_{d,t}^{ed}}, \forall i \in I, \tag{39}$$

where the higher λ_t indicates the more restoration of weighted loads and consequently the better performance of MG resilience enhancement. The designed reward function (39) is similar to the objective function (13), where the only difference is that the total weighted baseline loads are added in (39) to realize reward signals as unitless scalar values [21]. In Section II-C4, the objective of the problem is to maximize the weighted load restoration of the distribution network. However, directly using equation (13) as the reward function may raise serious convergence and optimality issues. This is because of the large fluctuations of weighted load restoration for different state conditions, possibly learning the unbalanced distributions of weight and bias values of the control policies [21]. To combat this, we scale the weighted load restoration and introduce the (unitless) resilience index $\lambda_t \in [0, 1]$ as the reward function.

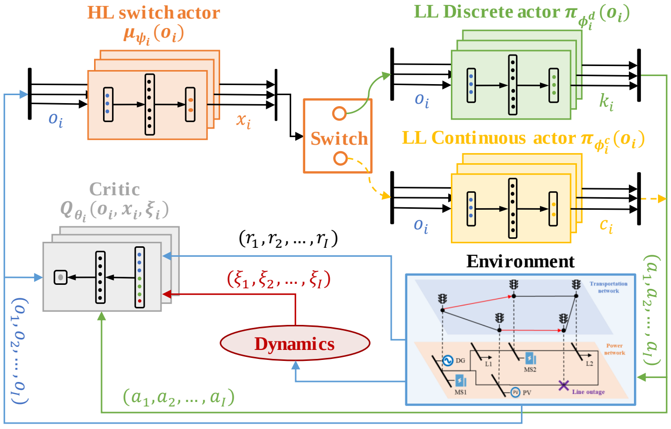

HL switch actor $\mu_{\psi_i}(o_i)$
o_i
x_i
Switch
LL Discrete actor $\pi_{\phi_i^d}(o_i)$
o_i
k_i
LL Continuous actor $\pi_{\phi_i^c}(o_i)$
o_i
c_i
Critic
$Q_{\theta_i}(o_i, x_i, \xi_i)$
$(r_1, r_2, \ldots, r_I)$
$(\xi_1, \xi_2, \ldots, \xi_I)$
Dynamics
Environment
Transportation network
Power network
DG
L1
MS2
L2
MS1
PV
Line outage
$(o_1, o_2, \ldots, o_I)$
$(a_1, a_2, \ldots, a_I)$
$(a_1, a_2, \ldots, a_I)$

Fig. 3. The structure of the proposed H2MAPPO method.

IV. Proposed MARL Method

In this section, a MARL method called H2MAPPO is proposed to solve the above Dec-POMDP, with the overall architecture depicted in Fig. 3. Specifically, H2MAPPO generates four practical implementation details that are insightful and crucial: 1) constructing a hierarchical architecture using a two-level framework [39] to choose between transport network routing and power network scheduling/repairing; 2) creating a hybrid policy [40] that can perform both discrete routing and continuous scheduling actions; 3) updating the MARL policy using MAPPO algorithm [41] that exhibits a stable learning performance, and is easy to implement, sample, and tune hyperparameters; 4) utilizing an embedded function to encapsulate system dynamics and approximate an abstracted state-value function, which can enhance the multi-agent training performance while providing privacy protection.

A. Learn Two-Level Hierarchies

Hierarchical reinforcement learning (HRL) refers to a type of RL method that can deal with several sub-policies working together in a hierarchical structure [42]. The two-level framework [39], one of the most common HRL techniques, is proposed as a temporal abstraction for RL actions, where the high-level (HL) action takes place over several time steps via the low-level (LL) actions. Specifically, for any RC or MPS agent i observing $o_{i,t}$ at time step t, an HL action is chosen using the HL policy $x_{i,t} = \mu(x|o) \to [0, 1] \in \mathcal{X}_i$ (e.g., when MESS i is parked to a MS at time step t, it has a $x_{i,t} = 80\%$ probability to choose charging/discharging actions in the power network and the left 20% probability to choose routing actions in the transport network, then the final HL action would be choosing charging/discharging behaviors in the power network, this is because final selection would be the one with the higher sampling probability). Afterwards, the LL policy $\pi(a|o)$ is utilized to compute the LL action $a_{i,t}$ (e.g., discharge MESS battery power to support load restorations). This process continues until the HL action switches to the transport network routing when probability $x_{i,t} < 50\%$. Similar as the vanilla RL, the reward over the two-level framework is given as $r_{i,t}$ in (39). Then, the objective of agent i over the proposed two-level framework within f time steps can be written as $R_i(o_{i,t}, x_{i,t}, o_{i,t+f}) = \mathbb{E}[\sum_{z=t}^{t+f} \gamma^{z-t} r_{i,z}]$. For each agent i, this process continues for T time steps, emitting a new trajectory of local observations, HL actions, LL actions, and rewards: $\tau_i = o_{i,1}, x_{i,1}, a_{i,1}, r_{i,1}, o_{i,2}, ..., r_{i,T}$ over $\mathcal{O}_i \times \mathcal{X}_i \times \mathcal{A}_i \times \mathcal{R} \rightarrow \mathbb{R}$.

In this perspective, we consider a two-level action policy as shown in Fig. 3, where agent i picks up an HL action $x_{i,t}$ according to its HL policy $\mu(x|o)$, and then computes the LL actions $a_{i,t}$ according to its LL policy $pi(a|o)$ at each time step t until the HL action $x_{i,t}$ switches to the other network environment. For each agent i, this process lasts for time steps T. In order to characterize the high-dimensional and continuous observation and action spaces of the HL and LL policies discussed above, an actor-critic architecture [21] is introduced for the hierarchical architecture. The actor module contains the HL policy $\mu(x|o)$ and the LL policy $\pi(a|o)$, while the critic module contains a state-value function $V(o, x)$ that specifies the expected value of selecting an HL selection $x_{i,t}$ in observation $o_{i,t}$. To deal with the complicated circumstances of the problem, deep neural networks (DNNs) are used as differentiable parameterized function approximators for both actor and critic modules. In detail, for each agent i, an actor network with parameters ψ_i is constructed for training the HL policy $\mu_{\psi_i}(x|o)$; another actor network with parameters ϕ_i is constructed for training the LL policy $\pi_{\phi_i}(a|o)$; and a critic network with parameters θ_i is constructed for training the state-value function $V_{\theta_i}(o, x)$.

B. Construct Hybrid Policy via MAPPO

After selecting the HL action $x_{i,t} = \mu_{\psi_i}(x|o)$ for either transport routing or power scheduling, each agent i at time step t should execute the LL actions $a_{i,t} = \pi_{\phi_i}(a|o)$ to the environment. Considering that the LL routing and scheduling actions of MEG and MESS agents are in discrete and continuous spaces respectively, a hybrid policy $a_{i,t} = \{k_{i,t}, c_{i,t}\} \in \mathcal{A}_i$ with two actor branches (networks) [40] is constructed to separately compute the discrete and continuous actions in the LL for MEG and MESS agents:[1]

$$k_{i,t} = \begin{cases} [a_{g,t}^{l,eg}] & \forall i = g \in I_{eg} \\ [a_{k,t}^{l,es}] & \forall i = k \in I_{es} \end{cases}, \tag{40}$$

$$c_{i,t} = \begin{cases} [a_{g,t}^{p,eg}] & \forall i = g \in I_{eg} \\ [a_{k,t}^{p,es}] & \forall i = k \in I_{es} \end{cases}. \tag{41}$$

In this case, MEG or MESS agent i will choose a discrete routing action $k_{i,t}$ when the HL action $x_{i,t}$ is switched to the transport network and choose a continuous scheduling action $c_{i,t}$ when the HL action $x_{i,t}$ is switched to the power network.

To model such action characteristics, we first generate a $\texttt{softmax}(\cdot)$ distribution for the discrete actor network parameterized by ϕ_i^d to output the corresponding probabilities for all potential routing behaviors, this categorical policy $\texttt{softmax}(o) = k_{i,t} = \pi_{\phi_i^d}(k|o)$ is then sampled for the optimal action $k_{i,t}$ in observation $o_{i,t}$. The softmax activation function takes in the local observation o and returns the probability scores of all possible discrete routing actions. In general, the equation of $\texttt{softmax}(\cdot)$ distribution is given as

$$\texttt{softmax}(o)_d = \frac{e^{o_d}}{\sum_{j=1}^{K} e^{o_j}}, \forall d \in 1, ..., K, \tag{42}$$

where K indicates the number of dimensions of LL discrete routing action $k_{i,t}$ in (40). In principle, function (42) applies the standard exponential function to each element o_d of the input local observation o and normalizes these values by dividing by the sum of all these exponentials; this normalization ensures that the sum of all elements' output probabilities is 1.

We then generate a Gaussian distribution $f(\cdot)$ for the continuous actor network parameterized by ϕ_i^c to output the corresponding mean and variance for all scheduling behaviors, the stochastic policy $f(o) = c_{i,t} = \pi_{\phi_i^c}(c|o)$ is then sampled for the optimal action $c_{i,t}$ in observation $o_{i,t}$. A Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(o) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{o-\mu}{\sigma}\right)^2} \tag{43}$$

where the parameter μ is the mean of the distribution, while the parameter σ is its standard deviation. The objective of the continuous actor network parameterized by ϕ_i^c is learning the parameters μ and σ in (43) given input local observation o.

The discrete policy $\pi_{\phi_i^d}$ and continuous policy $\pi_{\phi_i^c}$ are then updated independently using the MAPPO algorithm [41], which minimizes their clipped surrogate objective to restrict the policy update:

$$L_{i,t}^{\text{CLIP}}(\phi_i^d) = \hat{\mathbb{E}}_t\big[\min(\zeta_{i,t}^d \hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^d, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\big], \tag{44}$$

$$L_{i,t}^{\text{CLIP}}(\phi_i^c) = \hat{\mathbb{E}}_t\big[\min(\zeta_{i,t}^c \hat{A}_{i,t}, \text{clip}(\zeta_{i,t}^c, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t})\big], \tag{45}$$

where the former term in the operator $\min\{\cdot\}$ indicates the normal policy gradient while the later term in the operator $\min\{\cdot\}$ trims the policy gradient by clipping the probability ratio $\zeta_{i,t}^d, \zeta_{i,t}^c$ between $[1-\epsilon, 1+\epsilon]$. The hyperparameter $\epsilon \in [0, 1]$ is used to truncate the gradient update of the new policy from the old version. In other words, the advantage function $\hat{A}_{i,t}$ will be clipped if the probability ratio between the new and old policies goes beyond the range $[1 - epsilon, 1 + epsilon]$. In hybrid policy, the probability ratio $\zeta_{i,t}^d$ only takes into account the discrete policy, whereas the probability ratio $\zeta_{i,t}^c$ only takes into account the continuous policy. Specifically,

$$\zeta_{i,t}^d = \frac{\pi_{\phi_i^d}(k_{i,t}|o_{i,t})}{\pi_{\phi_i^d\text{old}}(k_{i,t}|o_{i,t})} \text{ and } \zeta_{i,t}^c = \frac{\pi_{\phi_i^c}(c_{i,t}|o_{i,t})}{\pi_{\phi_i^c\text{old}}(c_{i,t}|o_{i,t})}. \tag{46}$$

[1] It is noted here that both routing $a_{i,t}^{l,rc}$ and repairing $a_{i,t}^{r,rc}$ actions of RC agent are in discrete, we thus do not need the hybrid policy for RC agent, but adopt two categorical policies to generate $a_{i,t}^{l,rc}$ and $a_{i,t}^{r,rc}$, respectively.

Then, the HL discrete policy $\mu_{\psi_i}(x|o)$ can be updated similarly as the LL discrete policy $\pi_{\phi_i^d}(k|o)$ in (44), while the probability ratio of HL policy $\zeta_{i,t}^x$ can also be derived similarly as the LL discrete one $\zeta_{i,t}^d$ in (46).

In addition, the generalized advantage function $\hat{A}_{i,t}$ can be expressed as

$$\hat{A}_{i,t} = \delta_{i,t} + \gamma\delta_{i,t+1} + \cdots + \gamma^{T-t+1}\delta_{i,T-1}, \tag{47}$$

$$\delta_{i,t} = r_{i,t} + \gamma V_{\theta_i}(o_{1:I,t+1}, x_{1:I,t+1}) - V_{\theta_i}(o_{1:I,t}, x_{1:I,t}), \tag{48}$$

where $V_{\theta_i}(o,x)$ is the state-value function incorporating centralized training with the local observations $o_{1:I}$ and HL actions $x_{1:I}$ of all agents, which is approximated by a critic network parameterized by θ_i introduced in Section IV-A. It should be mentioned that providing all local information to the critic network can stabilize learning and foster coordinated behaviors for all local agents [41].

C. Abstract System Dynamics

However, the centralized critic network taking all agents' local information may raise problems. First, the shared information among all agents can destroy their privacy, since these decentralized MPSs and RCs are not willing to exchange their dispatch behaviors with each other [10]. Second, a POMDP may not be reduced to an MDP by concatenating all local information in centralised training since there may be crucial information (e.g., load shedding quantity) that is not noticed by any of the agents during training. This paper thus abstracts the system's global dynamics (e.g., load shedding quantity) via an embedded function ξ_i and approximates a new multi-agent joint state-value function as

$$V_{\theta_i}(o_{1:I}, x_{1:I}) = V_{\theta_i}(o_i, x_i, \xi_i), \tag{49}$$

which inputs individual agent's local observation o_i, HL action x_i, and embedded function ξ_i. Specifically, ξ_i indicates how much each agent i contributes to the system overall load restoration, which can be expressed as

$$\xi_i = \begin{cases} |P_j^{rc}| / \sum_{d\in D}(P_d^{ed} - P_d^{ls}) & \forall i = j \in I_{rc} \\ P_g^{eg} / \sum_{d\in D}(P_d^{ed} - P_d^{ls}) & \forall i = g \in I_{eg} \\ |P_k^{esd}| / \sum_{d\in D}(P_d^{ed} - P_d^{ls}) & \forall i = k \in I_{es} \end{cases}, \tag{50}$$

where $P_{i,t}^{rc}$ represents the power flow through the repaired line.

In this setting, ξ_i can be assumed to abstract the local observations of all other agents, e.g., $P_{i',t}^{ed}, P_{i',t}^{pv}, S_{i',t}^{ln}, \forall i' \in I(i')$, where $I(i')$ indicates the set of all other agents i' apart from i. Furthermore, ξ_i can reflect the status of agents supporting system resilience, i.e., the higher value of ξ_i denote the better performance of load restoration, and vice versa. As a consequence, each agent can make informative actions based on the abstracted knowledge ξ_i of local observations and actions from all other agents, albeit not directly obtaining their local information and dispatch behaviors, therefore safeguarding the privacy and enhancing scalability.[2]

D. Training Process

During the training process, H2MAPPO runs for all agents by their individual HL and LL policies $\mu_{\psi_i}(x|o), \pi_{\phi_i}(a|o)$ through T time steps, while collecting the trajectories τ_i (including the function ξ_i) from the interactions with the environment. Then, the agents can use the gathered trajectories to calculate the discounted reward-to-go $\hat{R}_{\iota,t} = \sum_{h=t}^{T} \gamma^{h-t} r_{\iota,h}$ and the advantage function $\hat{A}_{\iota,t}$ according to the abstracted state-value function $V_{\theta_i}(o_{\iota,t}, x_{\iota,t}, \xi_{\iota,t})$ for each trajectory ι and time step t, where a batch of trajectories are taken from the buffer $\mathcal{J} = \{\tau_\iota\} \sim \mathcal{F}$. Then, three actor networks are trained by maximising their objectives as follows:

$$\mathcal{L}(\psi_i) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta_{\iota,t}^x \hat{A}_{\iota,t}, \text{clip}(\zeta_{\iota,t}^x, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{51}$$

$$\mathcal{L}(\phi_i^d) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta_{\iota,t}^d \hat{A}_{\iota,t}, \text{clip}(\zeta_{\iota,t}^d, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{52}$$

$$\mathcal{L}(\phi_i^c) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(\zeta_{\iota,t}^c \hat{A}_{\iota,t}, \text{clip}(\zeta_{\iota,t}^c, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{53}$$

where J indicates the training batch size. The critic network is trained with the objective of minimizing the mean-squared error loss function:

$$\mathcal{L}(\theta_i) = \frac{1}{J}\sum_{\iota=1}^{J} \min\left(V_{\theta_i}(o_{\iota,t}, x_{\iota,t}, \xi_{\iota,t}) - \hat{R}_{\iota,t}\right)^2. \tag{54}$$

Given the above optimizations, the network weights of three actors and one critic can be updated as below:

$$\psi_i \leftarrow \psi_i + \alpha^{\psi} \nabla_{\psi_i} \mathcal{L}(\psi_i), \tag{55}$$

$$\phi_i^d \leftarrow \phi_i^d + \alpha^{\phi^d} \nabla_{\phi_i^d} \mathcal{L}(\phi_i^d), \tag{56}$$

$$\phi_i^c \leftarrow \phi_i^c + \alpha^{\phi^c} \nabla_{\phi_i^c} \mathcal{L}(\phi_i^c), \tag{57}$$

$$\theta_i \leftarrow \theta_i + \alpha^{\theta} \nabla_{\theta_i} \mathcal{L}(\theta_i), \tag{58}$$

where $\alpha^{\psi}, \alpha^{\phi^d}, \alpha^{\phi^d}, \alpha^{\theta}$ indicate the learning rates of the gradient ascent/descent algorithms for actor/critic networks. Finally, the pseudo-code of H2MAPPO is presented as below:

E. Test Process

The training process lasts for E episodes until the trained H2MAPPO method is being converged. Once in the test process, we firstly collect the weight parameters ψ_i of HL policy network as well as ϕ_i^d, ϕ_i^c of LL discrete and continuous policy networks respectively trained in Algorithm 1. The critic network is no longer required in the test process. For each time step in the test days F, each (RC, MEG, MESS) agent

[2] It should be noted that directly integrating function ξ_i into the current observation $o_{i,t}$ is unimplementable, since they can be only accessible once MGCC has solved the AC-OPF algorithm, which requires all agents' actions conditioned on the current local observations. However, the function ξ_i can be used in the critic training process, which is carried out after integrating them into the state-value function in (49).

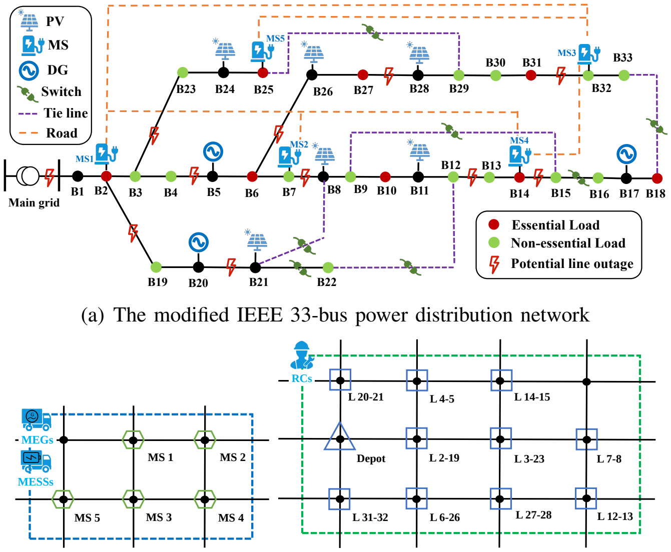
Algorithm 1 Training process of H2MAPPO for I agents

```
Initialize weights $\psi_i, \phi_i^d, \phi_i^c, \theta_i$ for actor and critic networks
Set learning rates $\alpha^{\psi}, \alpha^{\phi^d}, \alpha^{\phi^c}, \alpha^{\theta}$
for episode (i.e., day) $epi = 1$ to $E$ do
    Initialize the global state $s_0$ and local observation $o_{i,0}$
    For each agent $i$, sets an empty buffer $\mathcal{F} = \{\}$
    For each agent $i$, sets an empty trajectory $\tau_i = []$
    For each agent $i$, selects HL action $x_{i,0}$ in observing $o_{i,0}$
    for time step (i.e., 1 hour) $t = 1$ to $T$ do
        repeat
            for agent (i.e., RC, MEG, MESS) $i = 1$ to $I$ do
                Selects LL discrete action $a_{i,t} = k_{i,t}$ (if HL action is for transport) or continuous action $a_{i,t} = c_{i,t}$ (if HL action is for power)
            end for
            Execute all agents' actions $a_{1:I,t}$ to the environment, including both transport and power networks
            MGCC runs the AC-OPF algorithm once collecting all agents' dispatches (RCs' repairing decision, MEGs' power generation, and MESSs' charging and discharging power) and calculates the embedded function $\xi_{i,t}$ (if HL action is for power)
            for agent (i.e., RC, MEG, MESS) $i = 1$ to $I$ do
                Observes reward $r_{i,t}$ and next observation $o_{i,t+1}$
                Stores one sample experience to trajectory $\tau_i$ += $[o_{i,t}, x_{i,t}, a_{i,t}, r_{i,t}, \xi_{i,t}]$
                while time step $t \,\%\, J = 0$ do
                    Collects a set of trajectories $\tau_\iota$ from buffer $\mathcal{F}$, then computes advantage function $\hat{A}_{\iota,t}$ and discounted reward-to-go $\hat{R}_{\iota,t}$
                    Updates network weights $\psi, \phi^d, \phi^c, \theta$ in (55)-(58)
                end while
            end for
            Update local observation $o_{i,t} \leftarrow o_{i,t+1}$
        until HL action $x_{i,t}$ is switched in new observation $o_{i,t}$
        Update HL action $x_{i,t} \leftarrow x_{i,t+1}$
    end for
end for
```

i observes its current local observation $o_{i,t}$ and accordingly executes the HL action $x_{i,t}$ for environment switch as well as the LL discrete or continuous action $a_{i,t} = \{k_{i,t}, c_{i,t}\}$ for behaving in transport or power network. Those decisions are then mapped to the operation models of the coupled power-transport network (environment), transiting to the next state and time step. Each agent can also obtain its individual reward from the environment. Overall, the test process of the proposed H2MAPPO method is presented as below:

Algorithm 2 Test process of H2MAPPO for I agents

```
Load the weights $\psi_i, \phi_i^d, \phi_i^c$ trained by Algorithm 1
for test day = 1 : $F$ do
    Initialize the global state $s_0$ and local observation $o_{i,0}$
    For each agent $i$, selects HL action $x_{i,0}$ in observing $o_{i,0}$
    for time step (i.e., 1 hour) $t = 1$ to $T$ do
        repeat
            for agent (i.e., RC, MEG, MESS) $i = 1$ to $I$ do
                Selects LL discrete action $a_{i,t} = k_{i,t}$ (if HL action is for transport) or continuous action $a_{i,t} = c_{i,t}$ (if HL action is for power)
            end for
            Execute all agents' actions $a_{1:I,t}$ to the environment, including both transport and power networks
            MGCC runs the AC-OPF algorithm once collecting all agents' dispatches (RCs' repairing decision, MEGs' power generation, and MESSs' charging and discharging power) and calculates the load restoration $P_{d,t}^{ed}$
            for agent (i.e., RC, MEG, MESS) $i = 1$ to $I$ do
                Observes reward $r_{i,t}$ and next observation $o_{i,t+1}$
            end for
            Update local observation $o_{i,t} \leftarrow o_{i,t+1}$
        until HL action $x_{i,t}$ is switched in new observation $o_{i,t}$
        Update HL action $x_{i,t} \leftarrow x_{i,t+1}$
    end for
end for
```



(a) The modified IEEE 33-bus power distribution network

(b) MPSs' transport network (c) RCs' transport network

Fig. 4. The coupled power-transport network utilized for case studies: (a) the modified 33-bus power distribution network, (b) the transport network with MSs for MPSs, (c) the transport network with damaged components for RCs.

V. Case Studies

A. Experimental Setup

1) Network setup: To assess the effectiveness of the proposed MARL method in capturing realistic MPS and RC dispatch behaviors, a modified IEEE 33-bus power network, as shown in Fig. 4. The power network has 8 essential loads, 15 non-essential loads, 6 PVs, 3 DGs, and 5 MSs. Mobile resources deployed for load restoration include 1 MEG, 1 MESS and 1 RC. In the transport network, we assume that these resources can move to any candidate node through their routing characteristics, where detailed transport network structures between these candidate nodes (i.e., MSs for MPSs and damaged components for RCs) can be found in Fig. 4.

To capture the impact of extreme events, we assume that multiple line outages can happen in the power network, as depicted in Fig. 4. Specifically, the Monte Carlo sampling technique can be used to generate a manageable number of scenarios based on the fragility curve suggested in [7]. For each episode, a random outage scenario will be sampled from the fragility curve to represent the damaged conditions. Within the distribution network as shown in Fig. 4, the potential lines with higher damaged probabilities are easier to be selected into the outage scenario.

2) Data Descriptions: Case studies are conducted based on a real-world dataset from the Ausgrid [43]. The one-year

TABLE I
TECHNICAL PARAMETERS OF 3 DGS

DG	$\underline{P}^{dg}$ (kW)	$\overline{P}^{dg}$ (kW)	$\underline{Q}^{dg}$ (kVAR)	$\overline{Q}^{dg}$ (kVAR)
1	0	200	-67	100
2	0	300	-100	150
3	0	400	-133	200

TABLE II
TECHNICAL PARAMETERS OF MESS, MEG AND RC

RC		MEG		MESS	
$\overline{S}^{rc}$ (unit)	10	$\overline{P}^{eg}$ (kW)	150	$\overline{P}^{es}$ (kW)	100
RS^{rc} (unit)	[2,3]	$\overline{Q}^{eg}$(kVAR)	75	$\overline{E}^{es}$ (kWh)	400
RT^{rc} (h)	[1,4]	$\underline{Q}^{eg}$(kVAR)	-50	η^c/η^d (%)	90

PV generation and residential load data are collected, and then split into train (11 months) and test (1 month) sets for MARL method. To reflect the load distinction, 30% of loads are assumed to be essential featuring a high shedding cost at 2.5 £/kW, while the other 70% are non-essential and have a shedding cost at 1.5 £/kW. Tables I and II present the technical parameters of 3 static DGs and 3 mobile resources, respectively.

3) Benchmarks: The proposed H2MAPPO is compared with four benchmarks, including two MARL methods and two optimization methods: i) **IPPO**: each agent independently employs PPO , introducing a Gaussian policy with two dimensions, i.e., continuous scheduling action and discrete routing action by separating the continuous space into a finite set of segments; ii) **MAPPO**: based on IPPO, all local observations $o_{1:I}$ are concatenated by each agent to formulate its individual value function $V_i(o_{1:I})$; iii) **MPC**: MGCC employs a stochastic model predictive control (MPC) approach to solve a rolling optimization problem, which is constructed with the objective function (13) and constraints (1)-(12) and (15)-(32), assuming the perfect information of the power-transport network, mobile resource models, and all technical parameters; iv) **MILP**: MGCC employs a central deterministic mixed-integer linear programming (MILP) for the daily optimization problem with the objective function (13) and constraints (1)-(12) and (15)-(32), assuming the perfect knowledge of the system uncertainties.

4) Implementations: The training process of both actor and critic networks use the Adam optimizer with learning rates $\alpha^{\psi} = \alpha^{\phi^d} = \alpha^{\phi^c} = 10^{-4}$ and $\alpha^{\theta} = 10^{-3}$. The batch size $J = 24$ corresponds to 24 environment steps per episode. The discount rate $\gamma = 0.99$ and the clip rate $\epsilon = 0.2$. Multi-layer Perceptrons (MLPs) constructed by two hidden layers (128 and 64 units) with `ReLU` activation function are utilized for all networks. `Softmax` activation function is utilized to construct the categorical policy for both HL and LL discrete (routing and RC repair) actors. `Tanh` and `Softplus` activation functions are utilized to respectively construct a Gaussian policy with mean and standard deviation for the LL continuous (MPS scheduling) actor. Overall, the detailed specifications of actor and critic networks for three utilized MARL methods are presented in Table III. Both IPPO and MAPPO feature a single actor-critic network, the difference is that the input of critic network for IPPO is the individual local observation while the input of critic network for MAPPO is the all agents' local observations. For our proposed H2MAPPO, the network structure becomes more complicated. Specifically, there are four separate actor networks, indicating the HL actor for switch option; the first LL actor for routing decisions of both MPSs and RCs; the second LL actor for scheduling decisions of MPSs; and the third LL actor for repairing decisions of RCs. For all MARL methods, we run 5,000 episodes with the same 10 random seeds.

TABLE III
GENERAL SPECIFICATIONS OF THREE MARL METHODS.

Mechanism	Network	Structure
IPPO	Actor	linear(o_dim, 128) → ReLU()×2 → tanh+softplus(64, 2)
	Critic	linear(o_dim, 128) → ReLU()×2 → linear(64, 1)
MAPPO	Actor	linear(o_dim, 128) → ReLU()×2 → tanh+softplus(64, 2)
	Critic	linear(o_dim×$\|I\|$, 128)→ ReLU()×2 → linear(64, 1)
H2MAPPO	HL actor	linear(o_dim, 128) → ReLU()×2 → softmax(64, 2)
	LL actor (route)	linear(o_dim, 128) → ReLU()×2 → softmax(64, 4)
	LL actor (MPS)	linear(o_dim, 128) → ReLU()×2 → tanh+softplus(64, 1)
	LL actor (RC)	linear(o_dim, 128) → ReLU()×2 → softmax(64, 2)
	Critic	linear(o_dim+2, 128) → ReLU()×2 → linear(64, 1)

B. Performance Evaluation

In this subsection, the training and test performance of three investigated MARL methods are evaluated. Fig. 5 depicts the evolution of episodic reward over 5,000 training episodes, where the solid lines and the shaded areas respectively depict the moving average over 100 episodes and the oscillations of the original reward. Furthermore, their corresponding averaged episodic training time as well as the averaged number of episodes and averaged total training time required to reach convergence are collected in Table IV. Finally, we also collect the averaged resilience index (λ) and computation time over the 31 test days for three MARL and two optimization methods in Table V.

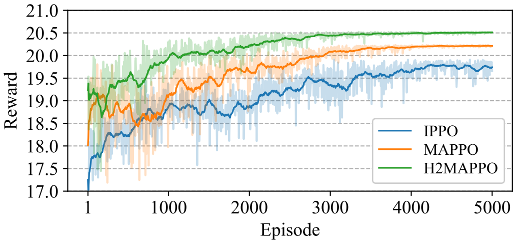


Fig. 5. Episodic reward over 5,000 episodes for different MARL methods.

TABLE IV
COMPUTATIONAL PERFORMANCE FOR DIFFERENT MARL METHODS

Method	IPPO	MAPPO	H2MAPPO
Episodic training time (s.)	1.83	2.14	2.65
Number of episodes (#)	5,000*	4,000	3,200
Total training time (hr.)	2.54*	2.38	2.36

* Fail to reach convergence within 5,000 episodes.

TABLE V
AVERAGED RESILIENCE INDEX AND COMPUTATION TIME OVER 31 TEST DAYS FOR DIFFERENT MARL AND OPTIMIZATION METHODS

Method	IPPO	MAPPO	H2MAPPO	MPC	MILP
Index-λ	19.56	20.15	21.34	19.93	22.12
Computation (sec.)	0.59	0.49	0.54	1028.93	76.21

The first observation we notice from Fig. 5 is that IPPO (blue) has the most unstable and oscillatory training performance, thereby obtaining the lowest reward level and failing to reach optimum. The independent learning method of IPPO, which concentrates on local information while disregarding the other agents, is considered to be the major cause of this instability issue, making the environment non-stationary. As a consequence, by concatenating all agents' information, MAPPO (orange) can effectively mitigate such non-stationarity and thus display better stability performance. However, MAPPO has inadequate policy quality due to the simple division of action space, which dramatically reduces its efficacy in handling scheduling actions in the power network, resulting in sub-optimum. Additionally, in the absence of a hierarchical architecture, each resource agent acquires routing and scheduling actions simultaneously, which may result in ineffective critic network learning, because there is always one meaningless action in the environment. In this case, the proposed H2MAPPO (green) can address the above issues by 1) utilizing embedded function ξ to learn system dynamics; 2) employing the hybrid policy to generate both continuous and discrete actions separately; and 3) introducing a hierarchical architecture that allows agents to adaptively switch to suitable environment status (power network or transport network).

We further assess the computational performance of three MARL methods during the training process. Table IV shows that IPPO has the shortest episodic training time (since it only requires training one actor network to compute both routing and scheduling/repairing actions, eliminating the need for hierarchical architecture), followed by MAPPO (since it takes all agents' local observations as the inputs of the centralized critic network), and H2MAPPO (since it needs to train four actor networks rather than the single actor network in IPPO and MAPPO). Additionally, we see that H2MAPPO (around 3,200 episodes) demonstrates a faster convergence rate than MAPPO (around 4,000 episodes). This is because the critic network incorporates an embedded function ξ_i that can stabilize the training performance and obtain a faster learning algorithm. Due to its instability issue, IPPO fails to reach convergence within 5,000 episodes. Finally, our proposed H2MAPPO (2.36 hrs) costs the similar computational time to MAPPO (2.38 hrs) but obtains a better policy quality (i.e., higher resilience level).

Regarding test performance in Table V, the proposed H2MAPPO achieves a near-to optimal performance (3.53% lower than MILP), and outperforms IPPO, MAPPO, and MPC in terms of the averaged resilience index over 31 test days by 9.10%, 5.91%, and 7.07%, respectively. On the other hand, all three MARL methods can be deployed in real-time around 0.5 sec., while the optimization-based MPC and MILP requires around 1000 sec. and 75 sec. averaged per day. It is worth noting that real-time control is important to the resilient MG operation problem due to the demand of a fast response time.

C. Analysis of Dispatch Behaviors and Switch Operations

After evaluating the MARL performance, this section validates the learned policy of H2MAPPO for dispatch behaviors of three resources, while the MG switch operations and load conditions are also involved. A scenario with 6 line outages (lines $4-5$, $14-15$, $2-19$, $3-23$, $6-26$ and $31-32$) is selected here. Additionally, serious traffic congestion mainly happens in the afternoon during the rush hours.

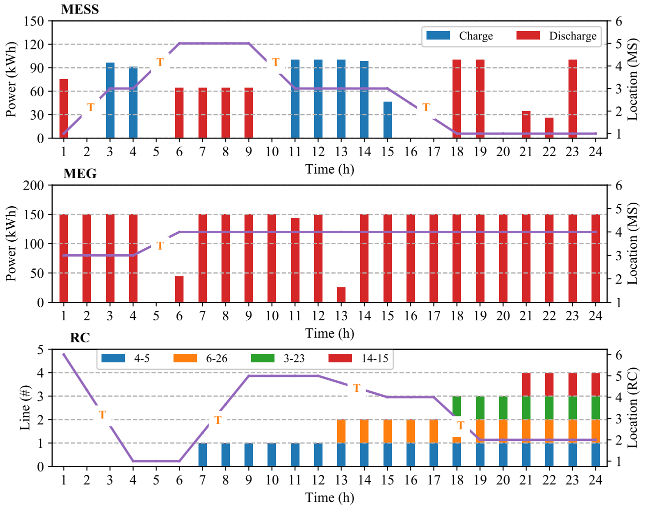


Fig. 6. Dispatch behaviors of MESS, MEG and RC.

TABLE VI
CONTRIBUTION OF MESS, MEG AND RC TO LOAD RESTORATION IN 33-BUS POWER NETWORK

Agent	MESS ($\lvert P^{esd} \rvert$)	MEG (P^{eg})	RC ($\lvert P^{rc} \rvert$)
Quantity (kWh)	692	3,211	6,363

1) Dispatch of MPSs and RCs: We first examine the dispatch behaviors of three MESS, MEG and RC agents, as depicted in Fig. 6. As for MESS in Fig. 6-(a), its routing behaviors are between MSs 1, 3 and 5. Specifically, the MESS chooses to discharge power at MSs 1 and 5 for demand supply, since both MS 1 at bus 2 and MS 5 at bus 25 connect with essential loads. Additionally, the discharging behaviors of MESS mainly occur at the periods of morning and night, when demand is relatively high. Now, let us look at the charging behaviors of MESS when it runs out of energy. The first charge occurs in the evening at MS 3 where MEG chooses to connect for power supply during the first few hours, as shown in Fig. 6-(b). Such phenomena also exhibits the coordination effect of MESS and MEG in both mobility and flexibility. The second charge occurs in the mid-day when free PV resources are abundant. Furthermore, the interesting results can be found that it takes MESS 2 hours (15:01-17:00) to travel from MS 3 to MS 1 in the afternoon while taking only 1 hour (1:01-2:00) from MS 1 to MS 3 in the morning. This is because the serious road congestion happening in the afternoon leads to another hour traveling time. On the other hand, MEG chooses to connect with MS 3 at bus 32 and MS 4 at bus 14 for power supply, since MS 4 is connected with essential load and one serious damage happens around bus 14. As for RC in Fig. 6-(c), it chooses to repair the damaged lines $4-5$, $6-26$, $3-23$ and $14-15$ sequentially. After these four lines are all repaired, RC has run out of its resources and is incapable of repairing more. It is also mentioned here the reason why

RC firstly repairs line $4-5$ is that repairing this line can restore the associated power flow, in which bus 2 is connected with essential load. In this case, there is no need for MEG to connect with MS 1 at bus 2 towards resilience enhancement. After analyzing the dispatch behaviours of three MESS, MEG and RC agents, we also summarize the overall contribution of each agent to the system resilience in Table VI. RC enhancing power flow contributes the most, followed by MEG and MESS. As such, it can be concluded that H2MAPPO successfully learns the reasonable dispatch behaviors for MESS, MEG and RC agents with the objective of providing resilience.

TABLE VII
SWITCH OPERATIONS FOR POWER NETWORK RECONFIGURATION

Time period (hr)	Switch Operations
1	close 8-21, 9-15, 12-22, 18-33, 25-29
13	open 25-29
18	open 26-27, 9-15 and 12-22
22	open 8-21

2) Switch Operations: Besides MESS, MEG and RC, switch operations that consider dynamic network reconfiguration can also help enhance resilience. It can be observed from Table VII that all five tie lines are closed to restore the obstructed power flow at the beginning of the day. But after RC gradually repairs the damaged lines, some switches open up to ensure the power network radiality. As a result, the resilience can be further enhanced via the smart network reconfiguration.

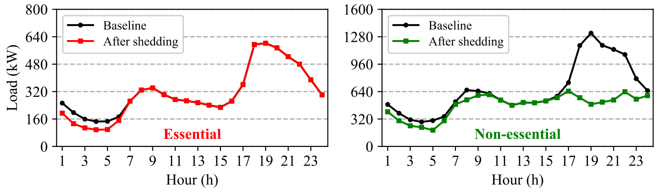


Fig. 7. Aggregated baseline and load after shedding in 33-bus system.

TABLE VIII
LOAD SHEDDING QUANTITY AND COST IN 33-BUS POWER NETWORK

Performance	Essential load	Non-essential load
Quantity (kWh)	291	4,217
Cost (£)	728	6,326

Finally, load conditions for both essential and non-essential types are compared in Fig. 7. Overall, the resilience enhancement for essential loads exhibits better performance than that for non-essential loads, respectively causing 291 kWh and 4,217 kWh total load shedding quantity, as compared in Table VIII. Thus, the system needs to pay serious cost for non-essential loads (6,326 £) compared to the essential loads (728 £).

D. Test Results in Modified IEEE 69-Bus Power Network

This section serves as a further demonstration of the proposed H2MAPPO on scalability. Thus, a modified IEEE 69-bus power network is introduced, which includes 7 DGs, 11 PVs and 10 MSs, while 4 MESSs, 4 MEGs and 4 RCs are

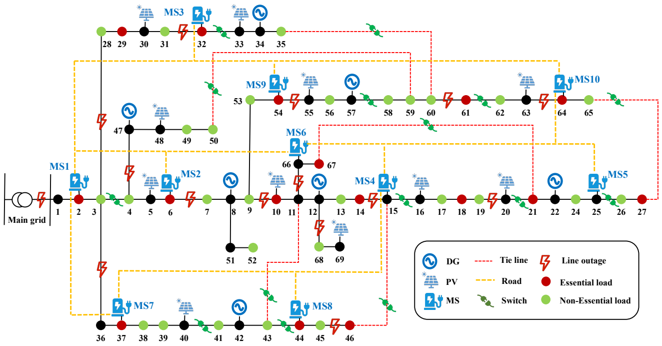


Fig. 8. The coupled 69-bus power-transport network.

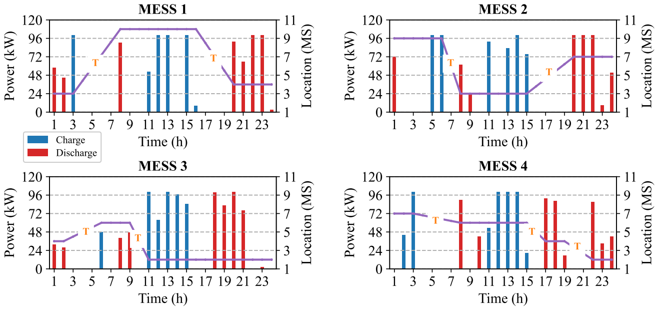


Fig. 9. Dispatch behaviors of MESSs in the modified 69-bus system.

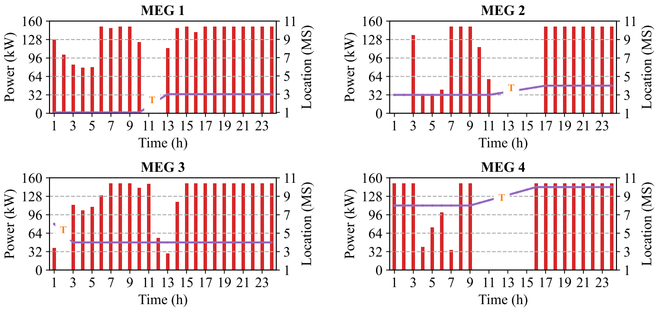


Fig. 10. Dispatch behaviors of MEGs in the modified 69-bus system.

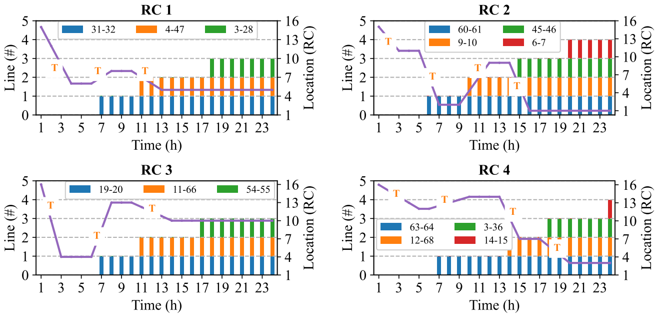


Fig. 11. Dispatch behaviors of RCs in the modified 69-bus system.

TABLE IX
LOAD SHEDDING QUANTITY AND COST IN 69-BUS POWER NETWORK

Performance	Essential load	Non-essential load
Quantity (kWh)	527	8,726
Cost (thous.£)	1,318	13,089

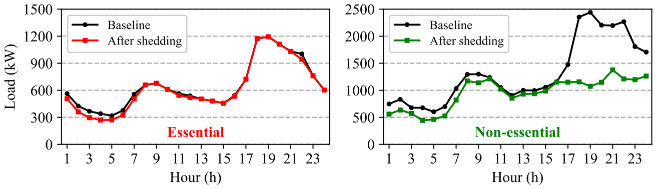


Fig. 12. Aggregated baseline and load after shedding in 69-bus system.

TABLE X
CONTRIBUTION OF MESSS, MEGS AND RCS TO LOAD RESTORATION IN 69-BUS POWER NETWORK

Agent	MESS ($\lvert P^{esd} \rvert$)	MEG (P^{eg})	RC ($\lvert P^{rc} \rvert$)
Quantity (kWh)	2,068	10,299	26,495

employed for load restoration. The detailed structure of the coupled power-transport network can be found in Fig. 8.

The dispatch behaviors of MESSs, MEGs, and RCs are illustrated in Figs. 9, 10, 11, respectively. Similar to the results in the 33-bus power network, MESSs and MEGs are coordinating with each other to provide power supply for the modified 69-bus power network towards timely load restoration. On one hand, 4 MESSs choose to charge power in the midday at MSs 2, 3, 6, 10 respectively due to their nearby high PV penetration, while discharging power in the evening at MSs connected with essential loads, e.g., MESS 1 at MS 4, MESS 2 at MS 7, MESS 3 at MS 2, and MESS 4 at MSs 2 and 4. On the other hand, MEGs move to the MSs connected with essential loads for power supply, e.g., MEG 1 at MSs 1 and 3, MEG 2 at MSs 3 and 4, MEG 3 at MS 4, and MEG 4 at MSs 8 and 10. Specifically, it can be found that MEG 3 moves to MS 4 and stays there all day, while MEG 2, and MESSs 1 and 4 are also connected with MS 4 in the evening when the load level reaches its peak. It is because the line outage occurring on line $45-46$ causes the isolation of the essential load at bus 46, while MPSs with black-start capabilities can energize bus 46 and provide power supply for this essential load via MS 4 and the smart switch operation on line $15-46$.

Furthermore, 4 RCs repair the damaged lines sequentially, where lines $31-32$, $60-61$, $19-20$, and $63-64$ are repaired in the first order. On one hand, repairing line $31-32$ can restore the power supply to the area around buses 28-31 including one essential load, since the only power source in this area is the grid-following PV at bus 30 without black-start capability. On the other hand, repairing lines $60-61$, $19-20$, and $63-64$ can restore the connections between the areas around buses 61-65 and 20-27 (including several essential loads) and the main power network. In particular, buses 61-63 can only be energized after line $60-61$ is repaired at 6:00, because of the lack of black-start capability in this area (grid-following PV at bus 63).

From the perspective of their coordination effect, it can be found that RCs firstly focus on repairing damaged components on the right part of the power network, while MPSs mainly choose to connect with MSs (e.g., MSs 1, 3, 4, and 8) on the left part of the power network. For instance, MEG 1 with black-start capability is connected with MS 1 to energize the area around buses 1-6 and provide power supply in the first 10 hours and then move to MS 3 when the line $4-47$ is repaired by RC 1 at 11:00. In this case, the DG unit at bus 47 is capable of providing power supply and black-start capability. Such coordination behaviors ensure that most buses in the power network can be energized quickly and enable fast load restoration.

Similar to the 33-bus system, the performance on providing resilience for essential loads is better than that for non-essential loads, as compared in Fig. 12 and Table IX. The cost for non-essential loads is thus much higher than that for essential loads. Furthermore, it can be observed in Table X that RCs enhancing power flow are also expected to contribute the most, followed by MEGs and MESSs. These results further validate the effectiveness of the proposed H2MAPPO in supporting system resilience for a large-scale system.

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel MARL method to address the coordinated dispatch problem of MPSs and RCs for MG load restoration. The proposed MARL method is characterized by a hierarchical architecture and a hybrid action domain including both discrete routing and continuous scheduling actions. The resilience-driven coordinated dispatch problem of MPSs and RCs is formulated as a Dec-POMDP, rendering a decentralized fashion and capturing the system dynamics of the coupled power-transport networks. Additionally, uncertainties related to renewables, demand, traffic volumes, and line outages are encompassed in the MARL training procedure. Experiment results based on two power networks (IEEE 33-bus and IEEE 69-bus) demonstrate the effectiveness of the coordinated dispatch of MPSs and RCs on restoring loads and enhancing resilience, while the outstanding performance of the proposed MARL method in optimality, stability, and scalability is testified, compared to the state-of-the-art MARL and optimization methods.

Future work aims at enhancing the studied problem from three directions. First, this paper focuses on the short-term daily load restoration problem. However, extreme events may have a long-term impact on the power system infrastructure. As a result, the first future extension is applying the proposed MARL method to a long-term load restoration problem that could last for several days/weeks. In this case, the episodic horizon will be expanded, e.g., 168 time steps for a 7-day episode with 1 hour per time step, while more efficient data sampling techniques could be applied to deal with the longer episodic horizon. Furthermore, these MPS and RC agents should be capable of learning multi-task policies within a single episode. For example, the re-filling process of resources or fuels that captures the influence of working time and resource availability can be considered in the RL training process by modifying the Dec-POMDP setup, e.g., appropriately adding a penalty term to the reward function to avoid the long-term working period. Second, this paper focuses on the routing and scheduling/repairing behaviors of MPSs and RCs, while their pre-allocation problem is not considered. However, the effective pre-allocation of these mobile and flexible sources can further improve system resilience. As a result, the second future extension is developing an optimal pre-allocation scheme for their initial positions and numbers of mobile sources towards system resilience enhancement. Third, this paper models the uncertainties of RL environment in terms of load profiles, PV generation, traffic congestion, and line outages, while the uncertainties of mobile sources themselves are not captured. However, in real-world applications, the mobile sources are also characterized by various uncertainties, e.g., the repair time of RCs, the charging losses of MESSs, and the fuel consumption of MEGs. As a result, the third future extension is taking the uncertainties of mobile sources into account and developing more realistic dispatch behaviors towards load restoration.

References

[1] Z. Bie, Y. Lin, G. Li, and F. Li, "Battling the extreme: A study on the power system resilience," *Proc. IEEE*, vol. 105, no. 7, pp. 1253–1266, Jul. 2017.

[2] T. Ding, M. Qu, Z. Wang, B. Chen, C. Chen, and M. Shahidehpour, "Power system resilience enhancement in typhoons using a three-stage day-ahead unit commitment," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2153–2164, May 2021.

[3] Z. Yang, P. Dehghanian, and M. Nazemi, "Seismic-resilient electric power distribution systems: Harnessing the mobility of power sources," *IEEE Trans. Ind. Appl.*, vol. 56, no. 3, pp. 2304–2313, May-Jun. 2020.

[4] Y. Wang, A. O. Rousis, and G. Strbac, "On microgrids and resilience: A comprehensive review on modeling and operational strategies," *Renew. Sust. Energ. Rev.*, vol. 134, p. 110313, Dec. 2020.

[5] A. Arif, S. Ma, Z. Wang, J. Wang, S. M. Ryan, and C. Chen, "Optimizing service restoration in distribution systems with uncertain repair time and demand," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6828–6838, Nov. 2018.

[6] J. Li, M. E. Khodayar, and M. R. Feizi, "Hybrid modeling based co-optimization of crew dispatch and distribution system restoration considering multiple uncertainties," *IEEE Syst. J.*, vol. 16, no. 1, pp. 1278–1288, Mar. 2022.

[7] 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.

[8] G. Zhang, F. Zhang, X. Zhang, Z. Wang, K. Meng, and Z. Y. Dong, "Mobile emergency generator planning in resilient distribution systems: A three-stage stochastic model with nonanticipativity constraints," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4847–4859, Nov. 2020.

[9] 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, p. 117921, 2022.

[10] S. Lei, C. Chen, H. Zhou, and Y. Hou, "Routing and scheduling of mobile power sources for distribution system resilience enhancement," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5650–5662, Sept. 2019.

[11] S. Lei, C. Chen, Y. Li, and Y. Hou, "Resilient disaster recovery logistics of distribution systems: Co-optimize service restoration with repair crew and mobile power source dispatch," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6187–6202, Nov. 2019.

[12] Z. Ye, C. Chen, B. Chen, and K. Wu, "Resilient service restoration for unbalanced distribution systems with distributed energy resources by leveraging mobile generators," *IEEE Trans. Industr. Inform.*, vol. 17, no. 2, pp. 1386–1396, Feb. 2021.

[13] T. Ding, Z. Wang, W. Jia, B. Chen, C. Chen, and M. Shahidehpour, "Multiperiod distribution system restoration with routing repair crews, mobile electric vehicles, and soft-open-point networked microgrids," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4795–4808, Nov. 2020.

[14] Q. Zhang, Z. Wang, S. Ma, and A. Arif, "Stochastic pre-event preparation for enhancing resilience of distribution systems," *Renew. Sust. Energ. Rev.*, vol. 152, p. 111636, Dec. 2021.

[15] Z. Wang, T. Ding, W. Jia, C. Mu, C. Huang, and J. P. Catalão, "Multi-period restoration model for integrated power-hydrogen systems considering transportation states," *IEEE Trans. Ind. Appl.*, vol. 58, no. 2, pp. 2694–2706, Mar.-Apr. 2021.

[16] T. Ding, Z. Wang, M. Qu, Z. Wang, and M. Shahidehpour, "A sequential black-start restoration model for resilient active distribution networks," *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 3133–3136, Jul. 2022.

[17] T. Ding, Y. Lin, G. Li, and Z. Bie, "A new model for resilient distribution systems by microgrids formation," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 4145–4147, Sept. 2017.

[18] T. Ding, Y. Lin, Z. Bie, and C. Chen, "A resilient microgrid formation strategy for load restoration considering master-slave distributed generators and topology reconfiguration," *Appl. Energy*, vol. 199, pp. 205–216, Aug. 2017.

[19] W. Li, Y. Li, C. Chen, Y. Tan, Y. Cao, M. Zhang, Y. Peng, and S. Chen, "A full decentralized multi-agent service restoration for distribution network with dgs," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1100–1111, Mar. 2019.

[20] P. Ge, F. Teng, C. Konstantinou, and S. Hu, "A resilience-oriented centralised-to-decentralised framework for networked microgrids management," *Appl. Energy*, vol. 308, p. 118234, Feb. 2022.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] J. C. Bedoya, Y. Wang, and C.-C. Liu, "Distribution system resilience under asynchronous information using deep reinforcement learning," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4235–4245, Sept. 2021.

[23] J. Zhao, F. Li, S. Mukherjee, and C. Sticht, "Deep reinforcement learning based model-free on-line dynamic multi-microgrid formation to enhance resilience," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 2557–2567, Jul. 2022.

[24] M. Kamruzzaman, J. Duan, D. Shi, and M. Benidris, "A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5525–5536, Nov. 2021.

[25] 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 *2020 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2020, pp. 1–5.

[26] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, "Multi-agent safe policy learning for power management of networked microgrids," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1048–1062, Mar. 2020.

[27] 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, p. 118575, Mar. 2022.

[28] Y. Wang, A. O. Rousis, and G. Strbac, "A three-level planning model for optimal sizing of networked microgrids considering a trade-off between resilience and cost," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5657–5669, Apr. 2021.

[29] Y. Wang, C. Chen, J. Wang, and R. Baldick, "Research on resilience of power systems under natural disasters—a review," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1604–1613, Feb. 2016.

[30] W. Yuanqing, Z. Wei, and L. Lianen, "Theory and application study of the road traffic impedance function," *J. Highway Transp. Res. Dev.*, vol. 21, no. 9, pp. 82–85, Nov. 2004.

[31] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Power Energy Mag.*, vol. 9, no. 4, pp. 101–102, Apr. 1989.

[32] Q. Zhang, Z. Ma, Y. Zhu, and Z. Wang, "A two-level simulation-assisted sequential distribution system restoration model with frequency dynamics constraints," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 3835–3846, Sept. 2021.

[33] F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.

[34] Q. Zhang, K. Dehghanpour, Z. Wang, and Q. Huang, "A learning-based power management method for networked microgrids under incomplete information," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1193–1204, Mar. 2019.

[35] G. Ruan, D. S. Kirschen, H. Zhong, Q. Xia, and C. Kang, "Estimating demand flexibility using siamese lstm neural networks," *IEEE Trans. Power Syst.*, vol. 37, no. 3, pp. 2360–2370, May 2022.

[36] J. Zhao, M. Mao, X. Zhao, and J. Zou, "A hybrid of deep reinforcement learning and local search for the vehicle routing problems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7208–7218, Nov. 2021.

[37] Y. Liang, Z. Ding, T. Zhao, and W.-J. Lee, "Real-time operation management for battery swapping-charging system via multi-agent deep reinforcement learning," *IEEE Trans. Smart Grid*, 2022.

[38] Y. Huang, Z. Ding, and W.-J. Lee, "Charging cost aware fleet management for shared on-demand green logistic system," *IEEE Internet Things J.*, 2022.

[39] D. Qiu, Y. Wang, M. Sun, and G. Strbac, "Multi-service provision for electric vehicles in power-transportation networks towards a low-carbon transition: A hierarchical and hybrid multi-agent reinforcement learning approach," *Appl. Energy*, vol. 313, p. 118790, May 2022.

[40] D. Qiu, Y. Wang, T. Zhang, M. Sun, and G. Strbac, "Hybrid multi-agent reinforcement learning for electric vehicle resilience control towards a low-carbon transition," *IEEE Trans. Industr. Inform.*, 2022.

[41] C. Yu, A. Velu, E. Vinitsky, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of mappo in cooperative, multi-agent games," *arXiv preprint arXiv:2103.01955*, 2021.

[42] R. S. Sutton, D. Precup, and S. Singh, "Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning," *Artif. Intell.*, vol. 112, no. 1-2, pp. 181–211, Aug. 1999.

[43] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray, "Residential load and rooftop pv generation: an australian distribution network dataset," *Int. J. Sustain. Energy*, vol. 36, no. 8, pp. 787–806, Oct. 2017.

Yi Wang received the Ph.D. degree from the Department of Electrical and Electronic Engineering at Imperial College London, U.K., in 2022. He is currently employed as a Research Associate in the Department of Electrical and Electronic Engineering at Imperial College London. His research interests include mathematical programming and learning approaches applied to the planning and operation of networked microgrids, the resilience enhancement of future power systems, and multi-energy system integration.

Dawei Qiu received the B.Eng. degree in Electrical and Electronic Engineering from Northumbria University, U.K., in 2014, the M.Sc. degree in Power System Engineering from University College London, U.K., in 2015, and the Ph.D. degree in Electrical Engineering Research from Imperial College London, U.K., in 2020. He is currently employed as a Research Associate in the Department of Electrical and Electronic Engineering at Imperial College London. His research focuses on the development and application of decentralized and market-based approaches to electricity market, peer-to-peer energy trading, multi-energy system integration, microgrid resilience control, and vehicle-to-grid flexibility. In particular, he has a strong background in game theoretic modelling and reinforcement learning approaches.

Fei Teng received the B.Eng in Electrical Engineering from Beihang University, China, in 2009, and the M.Sc. and Ph.D. degrees in Electrical Engineering from Imperial College London, U.K., in 2010 and 2015. Currently, he is a Lecturer in the Department of Electrical and Electronic Engineering, Imperial College London, U.K. His research focuses on cyber-physical modeling, optimization and data analytics of power systems.

Goran Strbac is a Professor of Energy Systems at Imperial College London, London, U.K. He led the development of novel advanced analysis approaches and methodologies that have been extensively used to inform industry, governments, and regulatory bodies about the role and value of emerging new technologies and systems in supporting cost effective evolution to smart low carbon future. He is currently the Director of the joint Imperial-Tsinghua Research Centre on Intelligent Power and Energy Systems, Leading Author in IPCC WG 3, Member of the European Technology and Innovation Platform for Smart Networks for the Energy Transition, and Member of the Joint EU Programme in Energy Systems Integration of the European Energy Research Alliance.