Reinforcement Learning with Enhanced Safety for Optimal Dispatch of Distributed Energy Resources in Active Distribution Networks

Xu Yang, Haotian Liu, Wenchuan Wu, *IEEE*, *Fellow*, Qi Wang, Peng Yu, Jiawei Xing, and Yuejiao Wang

***Abstract*—As numerous distributed energy resources (DERs) are integrated into the distribution networks, the optimal dispatch of DERs is more and more imperative to achieve transition to active distribution networks (ADNs). Since accurate models are usually unavailable in ADNs, an increasing number of reinforcement learning (RL) based methods have been proposed for the optimal dispatch problem. However, these RL based methods are typically formulated without safety guarantees, which hinders their application in real world. In this paper, we propose an RL based method called supervisor-projector-enhanced safe soft actor-critic (S3AC) for the optimal dispatch of DERs in ADNs, which not only minimizes the operational cost but also satisfies safety constraints during online execution. In the proposed S3AC, the data-driven supervisor and projector are pre-trained based on the historical data from supervisory control and data acquisition (SCADA) system, effectively providing enhanced safety for executed actions. Numerical studies on several IEEE test systems demonstrate the effectiveness and safety of the proposed S3AC.**

***Index Terms*—Reinforcement learning (RL), safety constraint, optimal dispatch, active distribution network (ADN), distributed energy resource (DER).**

I. Introduction

As large-scale distributed energy resources (DERs) are integrated into the distribution networks, the distribution networks are gradually transforming into active distribution networks (ADNs). Traditional passive control strategies can no longer effectively manage these new DERs and may cause severe security problems [1]-[3]. Therefore, to take full advantage of the DERs and ensure the safe operation of power system, an optimal dispatch is necessary in ADN operation.

Till now, the optimal dispatch of DERs including electric vehicles [4], energy storage devices [5], photovoltaic (PV) inverters [6], and wind farms [7], [8] is usually formulated as a P/Q coordinated optimization problem, which can be efficiently solved by some optimization methods. While these methods realize the optimal dispatch of DERs, most of them depend on complete and accurate models of ADNs, which are unaffordable to maintain for ADN operators [9], [10]. Therefore, to overcome model mismatch of the ADNs, deep reinforcement learning (RL) based methods have been widely adopted in power system operation, such as volt-var control [11], [12], optimal power flow [13], secondary control [14], and energy management [15].

Besides, as an efficient data-driven approach, the RL based method can also help solve the complex computational problems, which have been recently studied in some novel scenarios of the power system. For example, researchers in [16] design a joint electricity-carbon trading framework to reduce carbon emission through trading and demand response, in which an improved RL algorithm is utilized to develop the optimal trading strategy. In [17], researchers develop a bottom-up Energy Internet architecture, in which the operation of each microgrid is achieved by a curriculum learning enhanced RL based method. In [18], researchers propose an RL based method for the carbon-oriented optimal scheduling of electric vehicle aggregators in a complex distribution network, achieving lower cost and carbon emission with a higher efficiency.

However, when applying these RL based methods to real-world ADNs, safety becomes a critical concern. Because of the trial-and-error nature of RL algorithms and the lack of safety considerations, the RL agent may generate numerous unsafe actions and cause disastrous consequences. In order to satisfy the safety constraints, a number of constrained RL algorithms have been proposed recently, which can be roughly divided into four categories.

1) Prior knowledge based algorithms. These constrained RL algorithms utilize a partially known system model or other useful information to construct a safe action region, which

Manuscript received: November 15, 2023; revised: January 25, 2024; accepted: March 17, 2024. Date of CrossCheck: March 17, 2024. Date of online publication: April 18, 2024.

This work was supported in part by the National Key Research and Development Plan of China (No. 2022YFB2402900) and in part by the Science and Technology Project of State Grid Corporation of China "Key Techniques of Adaptive Grid Integration and Active Synchronization for Extremely High Penetration Distributed Photovoltaic Power Generation" (No. 52060023001T).

X. Yang, H. Liu, W. Wu (corresponding author), and Q. Wang are with the State Key Laboratory of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: yangxuthu@163.com; liuhaotian@tsinghua. edu. cn; wuwench@tsinghua. edu. cn; wangq19@mails. tsinghua. edu.cn).

P. Yu, J. Xing, and Y. Wang are with the State Grid Shandong Electric Power Company, Jinan 250000, China (e-mail: 167274738@qq. com; 573602466@qq. com; wangyuejiao2012@126.com).

DOI: 10.35833/MPCE.2023.000893

is updated periodically with the latest collected data. During the interaction with real-world entities, every executed action is strictly confined to this region [19], [20].

2) Reward shaping algorithms. In these algorithms, a penalty term for corresponding safety constraints is added to the reward received by the RL agent. When the safety constraints are violated, the RL agent will receive a negative reward. With the help of the added penalty term, the RL agent finally learns a relatively safe policy [21], [22].

3) Lagrangian algorithms. These algorithms can be regarded as an upgrade of the reward shaping ones, which formulate the Lagrangian function of the constrained optimization problem and iteratively update the policy of RL agent and Lagrangian multipliers to reach the optimal safe policy [23]-[25].

4) Constrained policy optimization. During the process of policy optimization, the policy of RL agent is restricted to a safe policy region using line search or projection. And the safe policy region is estimated based on the collected information when the safety constraints are violated [26], [27].

Some pioneering works also introduce the above constrained RL algorithms to develop data-driven safe control in power systems. For example, researchers in [28] incorporate Lagrangian algorithm with multi-agent RL based method to address the voltage safety constraints in ADNs. In [29], researchers utilize constrained policy optimization to achieve safe optimal operation in distribution networks. In [30], researchers add a barrier function as the penalty term in the reward received by the agent to keep it away from unsafe actions. In [31], researchers train an additional network to predict the frequency of the microgrid and a guidance network is utilized to promote safe learning.

Despite the great success of the above constrained RL algorithms, safety remains a critical concern for their further applications in real-world ADNs. Though these algorithms achieve the operational safety after online interactions, their safety during online interactions is not guaranteed. In addition, due to the absence of necessary “safety examination” and “safety modification” process, an unsafe action generated by the RL agent cannot be filtered out, which may cause unexpected consequences in real practice. Therefore, an action filter mechanism that provides enhanced safety during online interactions is indispensable to improve the applicability of current RL based methods.

To address the problems, we propose a safe RL based method called supervisor-projector-enhanced safe soft actor-critic (S3AC) for the optimal dispatch in ADNs, which ensures operational safety when interacting with real-world ADNs. As shown in Fig. 1, compared with a simple RL based method, the ADN controller in the proposed S3AC includes three main components: an RL agent, a supervisor, and a projector. The RL agent observes the states of the ADN and generates actions. The supervisor examines whether the generated actions are safe and the projector modifies the actions when necessary. Instead of directly executing generated actions on the real-world ADN, the supervisor and projector as the action guard successfully filter out unsafe ones, which provide enhanced safety for executed actions. To ensure the operational safety, the proposed S3AC consists of two stages: offline pre-training and online training & execution. In the offline stage, the supervisor and projector are pre-trained using the historical data from supervisory control and data acquisition (SCADA) system to formulate the action guard. Then, in the online stage, the ADN controller is transferred to interact with the real-world ADN and the performance of the RL agent is further enhanced.

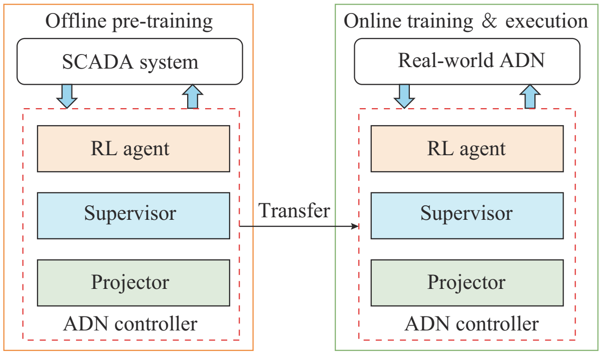

Fig. 1. Scheme of proposed S3AC.

In addition, due to poor data quality in ADN, considering possible outliers in historical measurements from SCADA system, the supervisor is formulated with a robust Gaussian process regression (GPR) and the neural network based projector is trained using the robust supervisor. A detailed description of the proposed S3AC will be presented in Section III. The unique contributions of this paper are summarized as follows.

1) A novel safe RL based method called S3AC is proposed to provide enhanced safety by introducing a supervisor and a projector. During online interactions, actions generated by the RL agent are first examined by the supervisor. If an action is determined unsafe by the supervisor, the projector projects it into a safe one with the minimum modification. This method efficiently filters out unsafe actions and ensures operational safety, which greatly enhances the applicability of current RL algorithms.

2) In order to address the model mismatch in ADNs, we leverage the historical data and propose a two-stage training mechanism under the scheme of the proposed S3AC. In the offline stage, the introduced supervisor and projector are first pre-trained using the historical data from SCADA system, which is completely data-driven. With only hundreds of instances, the supervisor achieves accurate predictions, based on which the neural network of projector successfully projects unsafe actions with the minimum modification.

3) Considering possible outliers in historical measurements, the data-driven supervisor is formulated with a robust GPR, which dynamically trims bad data in training samples. Compared with traditional GPR, this robust GPR effectively prevents interference from possible outliers, thus realizing a more accurate estimation of voltage magnitudes and branch power flow. The comprehensive experiments demonstrate the robustness of the robust GPR and safety of the proposed S3AC.

The remainder of this paper is organized as follows. Section II formulates the optimal dispatch problem in ADNs, and introduces the Markov decision process (MDP) and the soft actor-critic (SAC) RL algorithm used in this paper. Section III introduces the components and process of the proposed S3AC in detail. Section IV demonstrates the safety and effectiveness of the proposed S3AC and analyzes the results of different test cases. Finally, Section V states the conclusion and directions for future research.

II. Preliminaries

In this section, we first formulate the optimal dispatch problem of DERs in ADNs. Then, the settings of the MDP in this paper are explained. Lastly, we briefly introduce the SAC RL algorithm used in the proposed S3AC.

A. Formulation of Optimal Dispatch Problem

An ADN can be described as an undirected graph $\mathcal{G}(\mathcal{N},\mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ are the collections of nodes and branches, respectively. The DERs considered in this paper include micro gas turbines, PV inverters, and energy storage devices.

The power flow functions of ADN are shown as:

$$P_{i,t}=V_{i,t}\sum_{j\in\mathcal{N}}V_{j,t}\left(G_{ij}\cos\theta_{ij,t}+B_{ij}\sin\theta_{ij,t}\right)\quad\forall i\in\mathcal{N}\tag{1}$$

$$Q_{i,t}=V_{i,t}\sum_{j\in\mathcal{N}}V_{j,t}\left(-B_{ij}\cos\theta_{ij,t}+G_{ij}\sin\theta_{ij,t}\right)\quad\forall i\in\mathcal{N}\tag{2}$$

where $P_{i,t}$ and $Q_{i,t}$ are the active and reactive power injected into node i during time step t, respectively; $V_{i,t}$ and $V_{j,t}$ are the voltage magnitudes at nodes i and j during time step t, respectively; G_{ij} and B_{ij} are the real and imaginary parts of the corresponding element in the admittance matrix of ADN, respectively; and $\theta_{ij,t}$ is the voltage phase difference between nodes i and j during time step t.

Considering micro gas turbines, PV inverters, and energy storage devices, the power injection $P_{i,t}$ and $Q_{i,t}$ can also be expressed as:

$$P_{i,t}=P_{i,t}^{mt}+P_{i,t}^{PV}+P_{i,t}^{es}-P_{i,t}^{load}\tag{3}$$

$$Q_{i,t}=Q_{i,t}^{mt}+Q_{i,t}^{PV}+Q_{i,t}^{es}-Q_{i,t}^{load}\tag{4}$$

where the superscripts *mt*, *PV*, *es*, and *load* denote the micro gas turbines, PV inverters, energy storage devices, and power loads, respectively.

Without loss of generality, the DERs mentioned above are all controllable in this paper. We assume that the PV inverters operate in the maximum power point tracking (MPPT) mode, whose active power generation is time-varying and reactive power generation is under control. During time step t, the objective of the ADN operator is to minimize the operational cost by properly setting $P_{i,t}^{mt}$, $P_{i,t}^{es}$, $Q_{i,t}^{mt}$, $Q_{i,t}^{PV}$, and $Q_{i,t}^{es}$:

$$\min\sum_{t=0}^{T}\left(\sum_{i\in\mathcal{N}}C_i^{mt}(t)+\sum_{i\in\mathcal{N}}C_i^{es}(t)+C_0(t)\right)\tag{5}$$

$$C_i^{mt}(t)=\rho_i^{mt}P_{i,t}^{mt}\tag{6}$$

$$C_i^{es}(t)=\begin{cases}\rho_{i,dis}^{es}P_{i,t}^{es} & P_{i,t}^{es}\ge 0\\ -\rho_{i,ch}^{es}P_{i,t}^{es} & P_{i,t}^{es}<0\end{cases}\tag{7}$$

$$C_0(t)=\begin{cases}\rho_{buy,t}P_{0,t} & P_{0,t}\ge 0\\ \rho_{sell,t}P_{0,t} & P_{0,t}<0\end{cases}\tag{8}$$

where $C_i^{mt}(t)$ is the generation cost of the micro gas turbine at node i during time step t; $C_i^{es}(t)$ is the charging or discharging cost of the energy storage device at node i during time step t; $C_0(t)$ is the cost of buying electricity from the transmission network; ρ_i^{mt} is the cost coefficient of the micro gas turbine at node i; $\rho_{i,ch}^{es}$ and $\rho_{i,dis}^{es}$ are the charging and discharging cost coefficients of the energy storage device at node i, respectively; $\rho_{buy,t}$ and $\rho_{sell,t}$ are the prices of buying and selling electricity during time step t, respectively; and T is the length of an episode.

The constraints of micro gas turbines are shown as:

$$P_{i,\min}^{mt}\le P_{i,t}^{mt}\le P_{i,\max}^{mt}\quad\forall i\in\mathcal{N}\tag{9}$$

$$Q_{i,\min}^{mt}\le Q_{i,t}^{mt}\le Q_{i,\max}^{mt}\quad\forall i\in\mathcal{N}\tag{10}$$

$$-R_{i,down}\le P_{i,t}^{mt}-P_{i,t-1}^{mt}\le R_{i,up}\quad\forall i\in\mathcal{N}\tag{11}$$

where $R_{i,down}$ and $R_{i,up}$ are the maximum ramp-down and ramp-up rates, respectively; and the subscripts min and max represent the minimum and maximum values, respectively.

The constraints of PV inverters are shown as:

$$\left(P_{i,t}^{PV}\right)^2+\left(Q_{i,t}^{PV}\right)^2=\left(S_{i,t}^{PV}\right)^2\le\left(S_{i,\max}^{PV}\right)^2\quad\forall i\in\mathcal{N}\tag{12}$$

where $S_{i,\max}^{PV}$ is the installed capacity.

The constraints of energy storage devices are shown as:

$$P_{i,\min}^{es}\le P_{i,t}^{es}\le P_{i,\max}^{es}\quad\forall i\in\mathcal{N}\tag{13}$$

$$Q_{i,\min}^{es}\le Q_{i,t}^{es}\le Q_{i,\max}^{es}\quad\forall i\in\mathcal{N}\tag{14}$$

$$SOC_{i,\min}^{es}\le SOC_{i,t}^{es}\le SOC_{i,\max}^{es}\quad\forall i\in\mathcal{N}\tag{15}$$

$$SOC_{i,t}^{es}=\begin{cases}SOC_{i,t-1}^{es}-\dfrac{P_{i,t}^{es}\Delta t}{\eta} & P_{i,t}^{es}\ge 0,\ \forall i\in\mathcal{N}\\ SOC_{i,t-1}^{es}-\eta P_{i,t}^{es}\Delta t & P_{i,t}^{es}<0,\ \forall i\in\mathcal{N}\end{cases}\tag{16}$$

where $SOC_{i,t}^{es}$ is the state of charge of energy storage device at node i during time step t; Δt is the interval between two time steps; and η is the charging/discharging efficiency.

To ensure the safe operation of ADN, the voltage safety constraint (17) and branch capacity constraint (18) should be satisfied.

$$V_{\min}\le V_{i,t}\le V_{\max}\quad\forall i\in\mathcal{N}\tag{17}$$

$$\left(P_{ij,t}^{brch}\right)^2+\left(Q_{ij,t}^{brch}\right)^2=\left(S_{ij,t}^{brch}\right)^2\le\left(S_{\max}^{brch}\right)^2\quad\forall ij\in\mathcal{E}\tag{18}$$

where $P_{ij,t}^{brch}$ and $Q_{ij,t}^{brch}$ are the active and reactive power flows of branch ij during time step t, respectively; and $S_{ij,\max}^{brch}$ is the capacity of branch ij.

B. MDP

To apply RL algorithms in ADN operation, the optimal dispatch problem can be described as an MDP [32]. In MDP, an agent interacts with the environment, whose main components are defined by a tuple $(\mathcal{S},\mathcal{O},\mathcal{A},p,\mathcal{R},\gamma)$, where $\mathcal{S}$ is the state space of the environment; $\mathcal{O}$ and $\mathcal{A}$ are the observation space and action space of the agent, respectively; p is the state transition probability; $\mathcal{R}$ is the reward function; and $\gamma\in[0,1]$ is the discount factor for future rewards. It should be noted that the complete state space of the environment $\mathcal{S}$ is often inaccessible in reality. Instead, the observation space $\mathcal{O}$ obtained by the agent is a partial observation of $\mathcal{S}$. In this paper, we replace $o \in \mathcal{O}$ with $s \in \mathcal{S}$ and treat the unobserved states as noises according to the usual notation. During the interaction, the agent receives the observation of the environment $s_t \in \mathcal{S}$ and chooses an action $a_t \in \mathcal{A}$. After the action is executed on the environment, the environment transfers to the next state $s_{t+1} \in \mathcal{S}$ based on the unknown state transition probability $p: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, \infty)$. Then, a reward $r_t = \mathcal{R}(s_t, a_t, s_{t+1})$ is received by the agent, where the reward function $\mathcal{R}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ measures the performance of the agent in this transition.

The goal of the agent is to learn a policy π that optimizes its expected discounted reward $J(\pi)$:

$$\max_{\pi} \ J(\pi) = \mathbb{E}_{\tau \sim \pi}\left(\sum_{t=0}^{T} \gamma^t r_t\right) \tag{19}$$

where the policy π of agent is an action probability distribution in state s_t, i. e., $a_t \sim \pi(\cdot|s_t)$; and $\tau \sim \pi$ is the trajectory when the agent applies policy π to the environment.

To better illustrate the following SAC RL algorithm, we define an action value function $Q^{\pi}(s, a)$ in this section.

$$Q^{\pi}(s, a) = \mathbb{E}_{\tau \sim \pi}\left(\sum_{t=0}^{T} \gamma^t r_t | s_0 = s, a_0 = a\right) \tag{20}$$

where $Q^{\pi}(s, a)$ is the expected discounted reward after taking action a in state s with policy π.

Concerning a constrained MDP, when an agent interacts with the environment, the constraints can be divided into three parts: environment constraints, action constraints, and state constraints. The environment constraints are characteristics of the environment itself. The action constraints require the actions executed by the agent to be in a certain range. And the state constraints require the states of the environment to be safe enough. For the optimal dispatch problem described above, the power flow functions (1)-(4) are environment constraints, which are automatically satisfied by the physical system. The equipment constraints (6)-(16) are action constraints, which can be easily satisfied by appropriately setting the action space of the agent. The safety constraints (17) and (18) are state constraints, which are the main concern of the ADN operator and the focus of the proposed S3AC.

C. SAC

In this paper, the SAC RL algorithm is used for the design of RL agent [33]. It should be noted that any other RL algorithms can be easily implemented in our safe RL based method, we choose SAC here for its strong ability for exploration. Under the well-known actor-critic framework, SAC utilizes an actor network and a critic network to approximate its policy and the value function. The actor network π_θ, which consists of two subnetworks μ_θ and σ_θ, optimizes the policy with parameters θ. μ_θ and σ_θ share the same input layer and hidden layers, and output the mean value and standard deviation of the action distribution. The critic network Q_ϕ^π approximates the value function Q^π with parameters ϕ.

In SAC, when the RL agent interacts with the environment, it stores transitions (s_t, a_t, r_t, s_{t+1}) in a replay buffer $\mathcal{D}_{RL}$. Then, the neural networks of SAC are trained periodically with a batch of transition data $\mathcal{B}_{RL}$ randomly sampled from the buffer. The parameters ϕ of the critic network are optimized by minimizing the loss function $\mathcal{L}(\phi)$:

$$\mathcal{L}(\phi) = \frac{1}{|\mathcal{B}_{RL}|} \sum_{s_t, a_t, r_t, s_{t+1} \in \mathcal{B}_{RL}} \left(Q_\phi^\pi(s_t, a_t) - y_t\right)^2 \tag{21}$$

where y_t is the target value of $Q_\phi^\pi(s_t, a_t)$, which can be calculated using the Bellman equation:

$$y_t = r_t + \gamma\left(Q_{\phi'}^\pi(s_{t+1}, a_{t+1}) - \alpha \log \pi_\theta(a_{t+1}|s_{t+1})\right) \tag{22}$$

where $a_{t+1} \sim \pi_\theta(s_{t+1})$; $Q_{\phi'}^\pi$ is the target critic network introduced to stabilize the training process, whose parameters ϕ' are delayed parameters of ϕ and gradually updated using $\phi' \leftarrow \lambda_{\phi'} \phi + (1 - \lambda_{\phi'})\phi'$; and the term $-\log \pi_\theta(a_{t+1}|s_{t+1})$ is introduced by SAC to encourage the exploration of the RL agent, which effectively avoids overfitting to local optimal policies; and α is the corresponding coefficient of this term.

After the calculation of $\mathcal{L}(\phi)$, the parameters ϕ are updated using gradient descent:

$$\phi \leftarrow \phi - \lambda_\phi \nabla_\phi \mathcal{L}(\phi) \tag{23}$$

With the help of critic network, the parameters θ of the actor network are optimized by minimizing the loss function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}_{RL}|}\left(-Q_\phi^\pi(s_t, a_t) + \alpha \log \pi_\theta(a_t|s_t)\right) \tag{24}$$

where $a_t \sim \pi_\theta(s_t)$, and then parameters θ are also updated using gradient descent:

$$\theta \leftarrow \theta - \lambda_\theta \nabla_\theta \mathcal{L}(\theta) \tag{25}$$

In the equations above, $\lambda_{\phi'}$, λ_ϕ, and λ_θ are the learning rates of the corresponding parameters.

III. Methods

In this section, we innovate a safe RL based method, S3AC, for the optimal dispatch problem, which can effectively ensure the operational safety in ADN. This method consists of two stages: offline pre-training and online training & execution. In the offline stage, the supervisor and projector are pre-trained with historical data from the SCADA system. Then, in the online stage, the ADN controller interacts with the real-world ADN. The pre-trained supervisor and projector ensure the actions to be executed are safe enough.

A. MDP Setup for RL Agent

First, we formulate the optimal dispatch problem of ADN as an MDP, which can be efficiently optimized by the SAC. The definitions of state space, action space, and reward function are designed as follows.

1) State space: the state $\boldsymbol{s}_{RL}$ of the MDP is mainly based on the measurements in ADN, which is defined as $(\boldsymbol{P}^{mt}, \boldsymbol{P}^{PV}, \boldsymbol{P}^{es}, \boldsymbol{P}^{load}, \boldsymbol{Q}^{mt}, \boldsymbol{Q}^{PV}, \boldsymbol{Q}^{es}, \boldsymbol{Q}^{load}, \boldsymbol{V}, \boldsymbol{SOC}^{es})$, where each element in bold is the vector of corresponding variables.

2) Action space: the action $\boldsymbol{a}_{RL}$ of the MDP is based on controllable devices in ADN, and the upper and lower bounds of the action space satisfy equipment constraints (6)-(16). For DERs considered in this paper, the action $\boldsymbol{a}_{RL}$ is defined as $\left(\boldsymbol{P}^{mt}, \boldsymbol{P}^{es}, \boldsymbol{Q}^{mt}, \boldsymbol{Q}^{PV}, \boldsymbol{Q}^{es}\right)$.

3) Reward function: the reward function is based on the objective function and safety constraints of the optimal dispatch problem in ADN, which consists of the instant operational cost R_c, voltage violation rate VVR, and branch capacity violation rate SVR:

$$r_{RL,t} = -\beta_{RL,c} R_c(t) - \beta_{RL,V} \cdot VVR(t) - \beta_{RL,S} \cdot SVR(t) \tag{26}$$

$$R_c(t) = \sum_{i \in \mathcal{N}} C_i^{mt}(t) + \sum_{i \in \mathcal{N}} C_i^{es}(t) + C_0(t) \tag{27}$$

$$VVR(t) = \sum_{i \in \mathcal{N}} \left(\left[V_{i,t} - V_{\max}\right]_+^2 + \left[V_{\min} - V_{i,t}\right]_+^2 \right) \tag{28}$$

$$SVR(t) = \sum_{ij \in \mathcal{E}} \left[\left(S_{ij,t}^{brch} - S_{\max}^{brch}\right) \Big/ S_{\max}^{brch} \right]_+^2 \tag{29}$$

where $\beta_{RL,c}$, $\beta_{RL,V}$, and $\beta_{RL,S}$ are the corresponding coefficients; and $[\,\cdot\,]_+$ is the rectified linear unit function defined as $[x]_+ = \max(x, 0)$. If the safety constraints (17) and (18) are satisfied, we have $VVR(t) = 0$ and $SVR(t) = 0$.

B. Pre-training of Supervisor

As mentioned before, simply reward shaping as (26) cannot provide safety guarantees for actions to be executed, so the supervisor and projector are introduced in the proposed S3AC. The supervisor is pre-trained in the offline stage using historical data from SCADA system, whose role is to examine whether $\boldsymbol{a}_{RL}$ generated by the RL agent is safe enough. In this paper, the supervisor is formulated with a robust GPR to approximate the voltage magnitudes and branch power, which is robust to outliers in historical measurements.

In traditional GPR, the relationship between its input $\boldsymbol{x}$ and output y is modeled as a regression $y = f(\boldsymbol{x}) + \epsilon_n$, where $f(\boldsymbol{x})$ is the underlying model; and ϵ_n is a homoscedastic Gaussian noise. The objective of traditional GPR is to infer the underlying model based on a batch of data $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^{n}$, where n is the data size. Assume the underlying model $f(\boldsymbol{x})$ is a Gaussian process with zero mean value and kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$, and then $\boldsymbol{f} = \left\{f(\boldsymbol{x}_i)\right\}_{i=1}^{n}$ follows a multivariate Gaussian distribution $\boldsymbol{f} \sim N(\,\cdot\,|0, \boldsymbol{k})$, where $\boldsymbol{k}$ is the covariance matrix determined by $k_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

When given a new input $\boldsymbol{x}_*$, using the maximum likelihood estimation, the traditional GPR provides the posterior prediction of $f(\boldsymbol{x}_*)$, including the mean value $\hat{f}_* = \mathbb{E}\left(f(\boldsymbol{x}_*)\right)$ and variance $\sigma_*^2 = var\left(f(\boldsymbol{x}_*)\right)$. For a detailed description of the traditional GPR, please refer to [34].

Although traditional GPR achieves accurate approximation in theory, it can be severely biased when the data are contaminated, especially when there exist outliers in historical measurements, which is not reliable for safe operation. So, we utilize a robust GPR introduced in [35], which consists of shrinking, concentrating, and reweighting stages.

In the shrinking and concentrating stage, first train the standard GPR with the full batch of data $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^{n}$ and calculate the estimated mean value $\hat{f}_i$, variance σ_i^2, and normalized residual $r_i' = \left|y_i - \hat{f}_i\right| / \sigma_i$ for each point. Then, retrain the GPR using the δn points with the smallest residuals and update the estimated $\left\{\hat{f}_i, \sigma_i, r_i'\right\}$ for each point, where δ is the preserving fraction. The retrain and update step is repeated for $n_{sh} + n_{cc}$ times. The preserving fraction δ shrinks from 1 to a trimming parameter δ_1 in the first n_{sh} iterations and remains constant for the next n_{cc} iterations. Besides, because the variance of a $1 - \delta$ trimmed sample underestimates the actual variance of the underlying sample, the corrected normalized residual is expressed as $r_i = \left|y_i - \hat{f}_i\right| \Big/ \left(\sigma_i \sqrt{c}\right)$. Here, $c = \delta \Big/ F_{\chi_3^2}\left(\chi_{1,\delta}^2\right)$ is a consistency factor, where $F_{\chi_3^2}$ is the cumulative distribution function of the χ_3^2 distribution; and $\chi_{1,\delta}^2$ is the δ-quantile of the χ_1^2 distribution.

In the reweighting stage, remove the data points with $r_i^2 > \chi_{1,\delta_2}^2$ and retrain the GPR with the remaining samples, where δ_2 is a reweighting parameter. The process of the robust GPR is summarized in Algorithm 1.

Algorithm 1: process of robust GPR

```
Input: $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i=1}^{n}$, $\delta_1$, $\delta_2$, $n_{sh}$, and $n_{cc}$
Output: trimmed samples $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i \in I}$, trained GPR hyperparameters $\Theta$,
    and consistency factor $c$
for $j = 0$ to $n_{sh} + n_{cc}$ do
    if $j = 0$ then
        $I = \{1, 2, \ldots, n\}$
        $c = 1$
    else
        if $j \le n_{sh}$ then
            $\delta = 1 - (1 - \delta_1)\, j / (n_{sh} + 1)$
        else
            $\delta = \delta_1$
        end if
        $I = \left\{i \,\middle|\, r_i \le \delta - quantile(\boldsymbol{r})\right\}$
        $c = \delta \Big/ F_{\chi_3^2}\left(\chi_{1,\delta}^2\right)$
    end if
    $\Theta = gp\_optimize\left\{(\boldsymbol{x}_i, y_i)\right\}_{i \in I}$
    $\hat{\boldsymbol{f}}, \boldsymbol{\sigma}^2 = gp\_predict\left(\left\{\boldsymbol{x}_i\right\}_{i=1}^{n} \,\middle|\, \left\{(\boldsymbol{x}_i, y_i)\right\}_{i \in I}, \Theta\right)$
    $\boldsymbol{r} = \left|\boldsymbol{y} - \hat{\boldsymbol{f}}\right| \Big/ \left(\boldsymbol{\sigma}\sqrt{c}\right)$
end for
$I = \left\{i \,\middle|\, r_i^2 \le \chi_{1,\delta_2}^2\right\}$
$c = \delta_2 \Big/ F_{\chi_3^2}\left(\chi_{1,\delta_2}^2\right)$
$\Theta = gp\_optimize\left\{(\boldsymbol{x}_i, y_i)\right\}_{i \in I}$
Return $\left\{(\boldsymbol{x}_i, y_i)\right\}_{i \in I}$, $\Theta$, and $c$
```

In the proposed S3AC, we use the robust GPR to approximate the voltage magnitudes and branch power. The input of the supervisor is $\boldsymbol{x} = \left(\boldsymbol{P}^{PV}, \boldsymbol{P}^{load}, \boldsymbol{Q}^{load}, \boldsymbol{P}^{mt}, \boldsymbol{P}^{es}, \boldsymbol{Q}^{mt}, \boldsymbol{Q}^{PV}, \boldsymbol{Q}^{es}\right)$, and output of the supervisor includes two parts: $\boldsymbol{V}$ and $\boldsymbol{S}^{brch}$.

With the input $\boldsymbol{x}$, the estimated mean values and variances for voltage magnitudes of all nodes are denoted as $\hat{\boldsymbol{f}}_V(\boldsymbol{x})$ and $\boldsymbol{\sigma}_V^2(\boldsymbol{x})$, respectively; the estimated mean values and variances for branch power of all branches are denoted as $\hat{\boldsymbol{f}}_S(\boldsymbol{x})$ and $\boldsymbol{\sigma}_S^2(\boldsymbol{x})$, respectively. After the pre-training of the supervisor, it can be transferred online to examine whether $\boldsymbol{a}_{RL}$ is safe enough.

C. Pre-training of Projector

After the pre-training of the supervisor, the next step in the offline stage is the pre-training of the projector. As shown in Fig. 2, the projector dataset $\mathcal{D}_p$ is generated by the well pre-trained supervisor in the previous step.

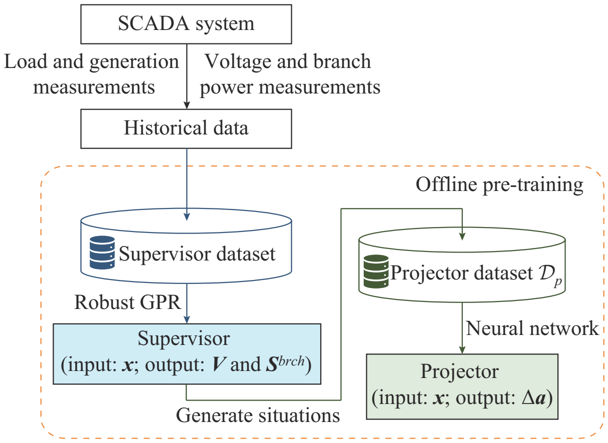
SCADA system
Load and generation measurements
Voltage and branch power measurements
Historical data
Offline pre-training
Supervisor dataset
Projector dataset $\mathcal{D}_p$
Robust GPR
Neural network
Supervisor (input: $\boldsymbol{x}$; output: $\boldsymbol{V}$ and $\boldsymbol{S}^{brch}$)
Projector (input: $\boldsymbol{x}$; output: $\Delta\boldsymbol{a}$)
Generate situations

Fig. 2. Offline pre-training stage of proposed S3AC.

The input of the projector is the same as that of the supervisor, i. e., $\boldsymbol{x}=\left(\boldsymbol{P}^{PV},\boldsymbol{P}^{load},\boldsymbol{Q}^{load},\boldsymbol{P}^{mt},\boldsymbol{P}^{es},\boldsymbol{Q}^{mt},\boldsymbol{Q}^{PV},\boldsymbol{Q}^{es}\right)$, which is the load and power generation in this ADN. To better illustrate the role of the projector, we divide the input $\boldsymbol{x}$ into two parts: the uncontrollable part $\boldsymbol{x}_{-a}=\left(\boldsymbol{P}^{PV},\boldsymbol{P}^{load},\boldsymbol{Q}^{load}\right)$ and the controllable part $\boldsymbol{a}_{RL}=\left(\boldsymbol{P}^{mt},\boldsymbol{P}^{es},\boldsymbol{Q}^{mt},\boldsymbol{Q}^{PV},\boldsymbol{Q}^{es}\right)$, i. e., $\boldsymbol{x}=\left(\boldsymbol{x}_{-a},\boldsymbol{a}_{RL}\right)$. The output of the projector is defined as $\Delta\boldsymbol{a}=\left(\Delta\boldsymbol{P}^{mt},\Delta\boldsymbol{P}^{es},\Delta\boldsymbol{Q}^{mt},\Delta\boldsymbol{Q}^{PV},\Delta\boldsymbol{Q}^{es}\right)$, which is modification to the controllable part. In the proposed S3AC, the role of the projector is to project the action generated by the RL agent $\boldsymbol{a}_{RL}$ to $\boldsymbol{a}_{RL}+\Delta\boldsymbol{a}$ with the minimal modification, so that the action after modification is safe enough for ADN operation. We denote $\boldsymbol{x}$ after modification as $\tilde{\boldsymbol{x}}$, i.e., $\tilde{\boldsymbol{x}}=\left(\boldsymbol{x}_{-a},\boldsymbol{a}_{RL}+\Delta\boldsymbol{a}\right)$. Therefore, the optimization problem for the projector can be formulated as:

$$\min \mathbb{E}_{x\sim\mathcal{D}_p}\|\Delta\boldsymbol{a}\|_2 \tag{30}$$

s.t.

$$\bar{f}_V(\tilde{\boldsymbol{x}})\leq V_{\max}\quad \forall\bar{f}_V(\tilde{\boldsymbol{x}})\in\bar{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}}) \tag{31}$$

$$\underline{f}_V(\tilde{\boldsymbol{x}})\geq V_{\min}\quad \forall\underline{f}_V(\tilde{\boldsymbol{x}})\in\underline{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}}) \tag{32}$$

$$\bar{f}_S(\tilde{\boldsymbol{x}})\leq S_{\max}^{brch}\quad \forall\bar{f}_S(\tilde{\boldsymbol{x}})\in\bar{\boldsymbol{f}}_S(\tilde{\boldsymbol{x}}) \tag{33}$$

$$\bar{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})=\hat{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})+d\boldsymbol{\sigma}_V(\tilde{\boldsymbol{x}}) \tag{34}$$

$$\underline{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})=\hat{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})-d\boldsymbol{\sigma}_V(\tilde{\boldsymbol{x}}) \tag{35}$$

$$\bar{\boldsymbol{f}}_S(\tilde{\boldsymbol{x}})=\hat{\boldsymbol{f}}_S(\tilde{\boldsymbol{x}})+d\boldsymbol{\sigma}_S(\tilde{\boldsymbol{x}}) \tag{36}$$

where d is the size of the confidence interval predicted by the supervisor. In this paper, we choose $d=3$. In the proposed S3AC, we formulate the projector with a neural network with parameters ξ. During its training process, the loss function $\mathcal{L}(\xi)$ is also calculated based on a batch $\mathcal{B}_p$ randomly sampled from $\mathcal{D}_p$:

$$\begin{aligned}\mathcal{L}(\xi)=\frac{1}{\left|\mathcal{B}_p\right|}\sum_{x\in\mathcal{B}_p}\Bigg[&\|\Delta a\|_2+\beta_{p,V}\sum_{\bar{f}_V(\tilde{\boldsymbol{x}})\in\bar{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})}\left[\bar{f}_V(\tilde{\boldsymbol{x}})-V_{\max}\right]_+^2+\\&\beta_{p,V}\sum_{\underline{f}_V(\tilde{\boldsymbol{x}})\in\underline{\boldsymbol{f}}_V(\tilde{\boldsymbol{x}})}\left[V_{\min}-\underline{f}_V(\tilde{\boldsymbol{x}})\right]_+^2+\\&\beta_{p,S}\sum_{\bar{f}_S(\tilde{\boldsymbol{x}})\in\bar{\boldsymbol{f}}_S(\tilde{\boldsymbol{x}})}\left[\left(\bar{f}_S(\tilde{\boldsymbol{x}})-S_{\max}^{brch}\right)/S_{\max}^{brch}\right]_+^2\Bigg]\end{aligned} \tag{37}$$

where $\beta_{p,V}$ and $\beta_{p,S}$ are the coefficients great enough to enforce the projector to satisfy these constraints.

Then, parameters ξ are updated using gradient descent:

$$\xi\leftarrow\xi-\lambda_\xi\nabla_\xi\mathcal{L}(\xi) \tag{38}$$

After the pre-training of projector, it can be transferred online to project $\boldsymbol{a}_{RL}$.

D. Online Training & Execution

After the offline pre-training, we can transfer the RL agent, the supervisor, and the projector to online training & execution. In the online stage, the ADN controller interacts with the real-world ADN. Because of the random exploration process, the RL agent could generate unsafe actions during this interaction, and this is where the added supervisor and projector start to function.

As shown in Fig. 3, in the online stage, we formulate the optimal dispatch of real-world ADN as an MDP. During the interaction, the RL agent first receives the observation of the ADN $\boldsymbol{s}_{RL}$ and generates an action $\boldsymbol{a}_{RL}$. Then, based on $\boldsymbol{s}_{RL}$ and $\boldsymbol{a}_{RL}$, the supervisor formulates its input $\boldsymbol{x}$ to examine whether $\boldsymbol{a}_{RL}$ is safe enough. If the estimated voltage magnitudes $\hat{\boldsymbol{f}}_V(\boldsymbol{x})$ and branch power $\hat{\boldsymbol{f}}_S(\boldsymbol{x})$ are within a safe range, $\boldsymbol{a}_{RL}$ is determined safe and executed directly on the real-world ADN. The real-world ADN transfers it to next state and generates a reward r_{RL} as feedback. The RL agent stores this transition information $\left(\boldsymbol{s}_{RL},\boldsymbol{a}_{RL},r_{RL}\right)$ in its replay buffer for SAC learning.

If the action is determined unsafe ($\hat{\boldsymbol{f}}_V(\boldsymbol{x})$ or $\hat{\boldsymbol{f}}_S(\boldsymbol{x})$ is out of safe range) by the supervisor, then $\boldsymbol{x}$ is sent to the projector, which gives $\Delta\boldsymbol{a}$ as output. The action executed on the real-world ADN is $\boldsymbol{a}_{RL}+\Delta\boldsymbol{a}$, which is safe enough. The real-world ADN transfers to the next state and generates a reward as feedback. Besides the transition information $\left(\boldsymbol{s}_{RL},\boldsymbol{a}_{RL}+\Delta\boldsymbol{a},r_{RL}\right)$, the RL agent also stores penalty information $\left(\boldsymbol{s}_{RL},\boldsymbol{a}_{RL},penalty\right)$ in the replay buffer, which indicates $\boldsymbol{a}_{RL}$ is an unsafe action under $\boldsymbol{s}_{RL}$, where *penalty* is a negative number with a large absolute value. With help of the penalty, the RL agent converges to a safe policy more quickly. If needed, the supervisor and projector can be updated periodically by the ADN operator in the control center with the latest collected data, so they can effectively approximate the current state of the system.

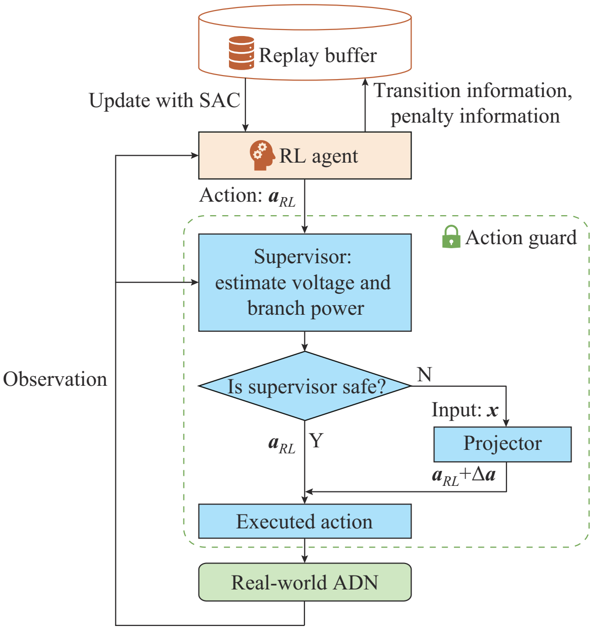

Fig. 3. Online training & execution stage of proposed S3AC.

Note that the proposed S3AC is not only useful for the operational safety in ADN optimal dispatch, but for a variety of safety constraints in different problems. In other model-free power system dispatch problems, the operators only need to find an efficient supervisor and an appropriate projector to help filter out unsafe actions.

IV. Numerical Study

In this section, numerical simulations are conducted on IEEE 33-bus [36], 69-bus [37], and 141-bus [38] test systems to validate the effectiveness and safety of the proposed S3AC. The steady-state distribution network RL environment is built under the scheme of the toolkit Gym [39] using the power flow functions. The branch capacity of the three systems are 3.2 MVA, 7.2 MVA, and 25.0 MVA, respectively, and voltage limitations are [0.95, 1.05]. Specific parameters of DERs in the three test systems are listed in Table I. More detailed configurations concerning the test systems are shown in the Supplementary Material.

A. Evaluation of Supervisor and Projector

In the offline stage of the proposed S3AC, the robust GPR based supervisor is trained with $\delta_1=0.5, \delta_2=0.975, n_{sh}=2, n_{cc}=2$. The numbers of instances for training supervisors in the three test systems are 800, 1000, and 1600, respectively. To verify the robustness of the robust GPR, we add normal distribution noise in half of the historical data to represent possible measurement errors or outliers. We compare robust GPR with traditional GPR, which is trained with the contaminated data without sample trimming.

TABLE I
Parameters of DERs in Three IEEE Test Systems

IEEE test system	DER type	Number	Parameter	Connected bus No.
33-bus	Micro gas turbine	2	Active power: 1.5 MW Reactive power: 1.5 Mvar	18, 33
	PV inverter	2	Capacity: 0.85 MVA Active power: 0.75 MW	22, 25
	Energy storage device	2	Active power: 0.3 MW Reactive power: 0.3 Mvar	21, 24
69-bus	Micro gas turbine	2	Active power: 3.0 MW Reactive power: 3.0 Mvar	18, 58
	PV inverter	2	Capacity: 1.8 MVA Active power: 1.5 MW	35, 46
	Energy storage device	2	Active power: 0.6 MW Reactive power: 0.6 Mvar	34, 45
141-bus	Micro gas turbine	2	Active power: 6.0 MW Reactive power: 6.0 Mvar	23, 55
	PV inverter	4	Capacity: 4.8 MVA Active power: 4.0 MW	49, 89, 116, 123
	Energy storage device	1	Active power: 1.2 MW Reactive power: 1.2 Mvar	91

The average and maximum normalized absolute errors of the predicted voltage magnitudes $\left|\hat{\boldsymbol{f}}_V(\boldsymbol{x})-\boldsymbol{V}\right|$ and branch power $\left|\hat{\boldsymbol{f}}_S(\boldsymbol{x})-\boldsymbol{S}^{brch}\right|/S_{\max}^{brch}$ are shown in Table II. As shown in Table II, the supervisor only needs hundreds of instances to achieve accurate predictions. And the maximum normalized absolute errors of predicted voltage magnitudes and branch power flow are small enough to generate the projector dataset and examine whether an action generated by RL agent is safe for ADN operation. In addition, compared with traditional GPR, the robust GPR effectively avoids the interference caused by outliers, which achieves higher approximation accuracy.

TABLE II
Performance of Supervisor in Offline Stage

IEEE test system	Parameter	Robust GPR		Traditional GPR	
		Mean	Maximal	Mean	Maximal
33-bus	$\left\|\hat{\boldsymbol{f}}_V(\boldsymbol{x})-\boldsymbol{V}\right\|$	3.17×10^{-4}	3.93×10^{-3}	4.05×10^{-3}	4.41×10^{-2}
	$\left\|\hat{\boldsymbol{f}}_S(\boldsymbol{x})-\boldsymbol{S}^{brch}\right\|/S_{\max}^{brch}$	1.04×10^{-2}	1.05×10^{-1}	3.30×10^{-2}	4.28×10^{-1}
69-bus	$\left\|\hat{\boldsymbol{f}}_V(\boldsymbol{x})-\boldsymbol{V}\right\|$	3.12×10^{-4}	4.06×10^{-3}	3.57×10^{-3}	3.19×10^{-2}
	$\left\|\hat{\boldsymbol{f}}_S(\boldsymbol{x})-\boldsymbol{S}^{brch}\right\|/S_{\max}^{brch}$	7.53×10^{-3}	1.13×10^{-1}	1.83×10^{-2}	4.13×10^{-1}
141-bus	$\left\|\hat{\boldsymbol{f}}_V(\boldsymbol{x})-\boldsymbol{V}\right\|$	4.52×10^{-4}	7.07×10^{-3}	3.86×10^{-3}	3.35×10^{-2}
	$\left\|\hat{\boldsymbol{f}}_S(\boldsymbol{x})-\boldsymbol{S}^{brch}\right\|/S_{\max}^{brch}$	3.59×10^{-3}	9.51×10^{-2}	8.94×10^{-3}	3.16×10^{-1}

After the pre-training of supervisor, the projector is then pre-trained with the dataset generated by the supervisor. In this paper, the neural network of the projector is trained with 2.0×10^5 instances generated by the supervisor. After its convergence, we randomly sample 100 load and power generation situations to test the projector. The average norm of the action modification $\|\Delta\boldsymbol{a}\|_2$, *VVR*, and *SVR* calculated by supervisor of the state after modification in these 100 situations are listed in Table III.

TABLE III
PERFORMANCE OF PROJECTOR IN OFFLINE STAGE

Test system	$\lVert \Delta \boldsymbol{a} \rVert_2$	*VVR*	*SVR*
33-bus	0.607	9.60×10^{-8}	0
69-bus	0.749	6.65×10^{-7}	0
141-bus	1.040	1.36×10^{-6}	0

From the results listed in Table III, we observe that in the final stage of the pre-training, all of the projectors converge to a stable performance, which are able to use a relatively small modification on the action to ensure the operational safety.

B. Setup of Proposed and Benchmark Methods

To verify the effectiveness of the proposed S3AC, we first formulate the SAC Lagrangian (SAC-L) without supervisor and projector as a state-of-the-art constrained RL baseline, which iteratively updates the policy of RL agent and Lagrangian multipliers to reach the safe optimal policy. In addition, to test the robustness of the robust GPR based supervisor, we formulate S3AC with traditional GPR, i. e., S3AC non-robust (S3AC-NR), whose supervisor is trained with the contaminated data without sample trimming. The hyperparameters of RL agents are listed in Table IV. A mixed-integer second-order cone programming (MISOCP) based on DistFlow [40] of the ADN is also implemented for comparison, which could be considered a theoretically optimal result. The details of MISOCP are presented in the Supplementary Material.

TABLE IV
HYPERPARAMETERS OF RL AGENTS

Hyperparameter	Value
Optimizer	Adam
Non-linearity	ELU
Replay buffer size	8000
Batch size	256
Number of hidden layers	2
Number of hidden units	256
Number of episode steps	96
γ	0.99
$\lambda_{\phi'}$	5.0×10^{-3}
λ_{ϕ}	1.0×10^{-3}
λ_{θ}	1.0×10^{-4}
α	2.0×10^{-2}

All of the RL agents are implemented in Python with the deep learning framework PyTorch. The MISOCP utilizes the commercial solver Gurobi. Experiments are run on a computer with a 2.3 GHz Intel Core i7-10875H CPU and 16 GB RAM. Due to the stochastic property of RL algorithms, we use three independent random seeds for each group of experiments.

C. Evaluation of Proposed S3AC

After the pre-training stage and algorithm setup, the ADN controller can be transferred to interact with the real-world ADN. We select PV and load data of one test day to test the training effect of the proposed S3AC, S3AC-NR, SAC-L, and MISOCP, whose profiles are also depicted in the Supplementary Material. Figures 4-6 display how these methods perform on this test day as the online training progresses in IEEE 33-bus system, 69-bus system, and 141-bus system, respectively, including total operational cost of the day and step average *VVR*. The mean values and standard deviations (std.) are presented as solid lines and filled areas, respectively.

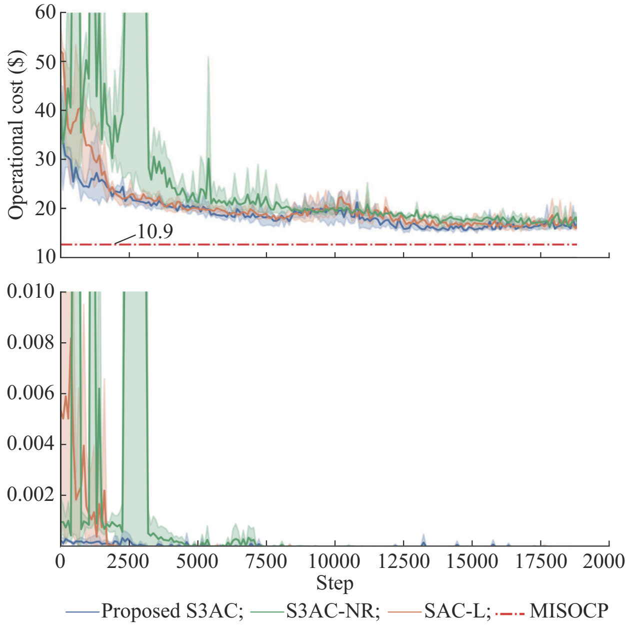

Fig. 4. Test results of proposed S3AC and benchmark methods in IEEE 33-bus system.

The total operational cost and step average *VVR* on this test day after convergence are listed in Table V. The maximum violation of safety index (step average *SVR* and step average *VVR*) on this test day during online training is listed in Table VI. The best performance among these methods is marked in bold.

First, by comparing the convergence rate of these methods, it can be observed from Figs. 4-6 that the proposed S3AC converges as quickly as SAC-L, which indicates the introduced supervisor and projector won't deteriorate the normal training process of the RL agent.

Meanwhile, for their final performance, both the proposed S3AC and SAC-L converge to a relatively low operational cost near the theoretically optimal result from MISOCP, which verifies the effectiveness of the model-free RL algorithms on the optimal dispatch problems in ADNs. With appropriate RL algorithm and MDP settings, the RL agent could achieve excellent performance comparable with traditional model based methods.

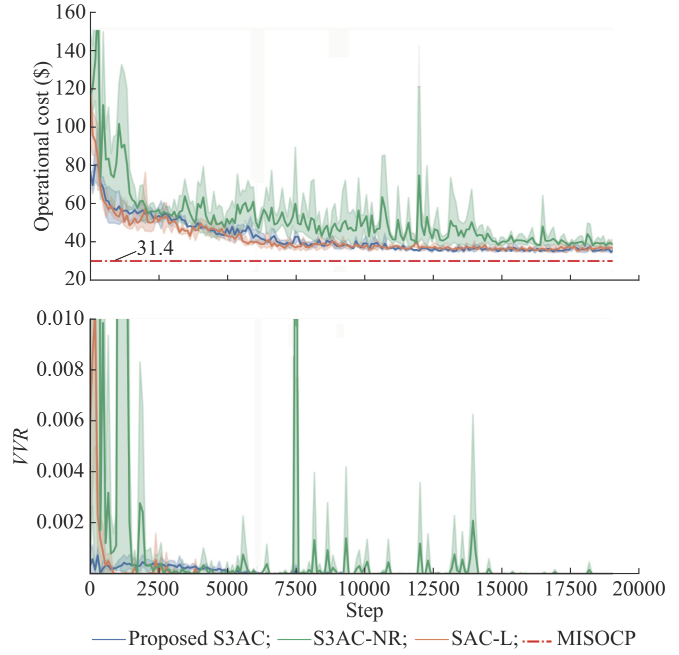

Fig. 5. Test results of proposed S3AC and benchmark methods in IEEE 69-bus system.

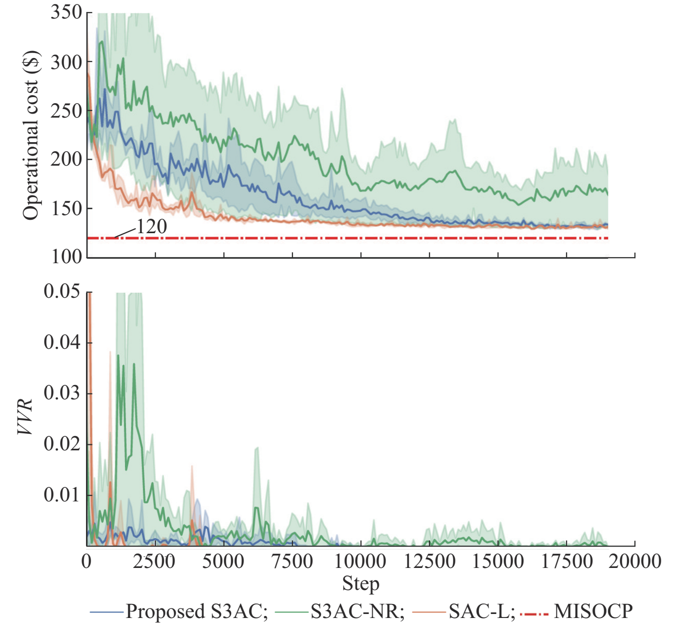

Fig. 6. Test results of proposed S3AC and benchmark methods in IEEE 141-bus system.

As for S3AC-NR in our experiments, its supervisor is formulated using traditional GPR, which is trained with the contaminated data without sample trimming. As demonstrated above, this supervisor cannot accurately approximate the voltage and branch power of the ADN, based on which the projector cannot give suitable action modifications either. These improper information from the supervisor and inappropriate modifications from the projector finally affect the normal interaction of the RL agent and worsen its training process. As can be observed from Figs. 4-6, the training process for S3AC-NR is much more unstable, and its final performance also degrades.

TABLE V
ONLINE PERFORMANCE AFTER CONVERGENCE OF PROPOSED S3AC AND BENCHMARK METHODS

Test system	Method	Operational cost ($)		*VVR*	
		Mean	Std.	Mean	Std.
33-bus	S3AC	15.4	0.895	**0**	**0**
	S3AC-NR	**15.0**	0.877	**0**	**0**
	SAC-L	15.3	**0.548**	4.98×10^{-6}	7.05×10^{-6}
	MISOCP	10.9		0	
69-bus	S3AC	**37.5**	0.378	**0**	**0**
	S3AC-NR	39.4	**0.166**	**0**	**0**
	SAC-L	37.5	0.894	**0**	**0**
	MISOCP	31.4		0	
141-bus	S3AC	134	2.781	**0**	**0**
	S3AC-NR	161	20.615	9.36×10^{-5}	1.12×10^{-4}
	SAC-L	**132**	**1.584**	**0**	**0**
	MISOCP	120		0	

TABLE VI
MAXIMAL VIOLATIONS DURING ONLINE INTERATION OF PROPOSED S3AC AND BENCHMARK METHODS

Test system	Method	*SVR*		*VVR*	
		Mean	Std.	Mean	Std.
33-bus	S3AC	$\mathbf{6.05\times10^{-3}}$	$\mathbf{7.48\times10^{-3}}$	$\mathbf{4.57\times10^{-4}}$	$\mathbf{1.17\times10^{-4}}$
	S3AC-NR	6.81×10^{-1}	5.95×10^{-1}	4.12×10^{-2}	3.09×10^{-2}
	SAC-L	2.38×10^{-1}	1.45×10^{-1}	1.14×10^{-2}	8.45×10^{-3}
69-bus	S3AC	$\mathbf{1.53\times10^{-5}}$	$\mathbf{8.68\times10^{-6}}$	$\mathbf{9.67\times10^{-4}}$	$\mathbf{1.55\times10^{-4}}$
	S3AC-NR	5.43×10^{-1}	2.33×10^{-1}	9.53×10^{-2}	4.28×10^{-2}
	SAC-L	1.23×10^{-1}	1.05×10^{-1}	1.07×10^{-2}	4.30×10^{-3}
141-bus	S3AC	$\mathbf{1.26\times10^{-5}}$	$\mathbf{1.77\times10^{-5}}$	$\mathbf{9.25\times10^{-3}}$	$\mathbf{1.47\times10^{-3}}$
	S3AC-NR	2.84×10^{-3}	2.89×10^{-3}	7.20×10^{-2}	3.01×10^{-2}
	SAC-L	1.87×10^{-2}	1.23×10^{-2}	8.33×10^{-2}	3.32×10^{-2}

When we apply model-free RL algorithms to real-world ADNs, the safety is a critical concern. It can be observed from Figs. 4-6 that SAC-L finally converges to a safe policy, which verifies the effectiveness of this constrained RL algorithm. However, at the beginning of the online interaction, unsafe actions are still generated and cause significant voltage violations. In real-world ADNs, these violations may lead to serious consequences.

However, with the help of the introduced supervisor and projector, the proposed S3AC becomes the safest one from the beginning, which keeps *VVR* at the lowest level, greatly improving the applicability of RL algorithms. Additionally, for S3AC-NR, since the supervisor and projector cannot provide accurate estimation and correct modification, its performance on safety index is even worse than SAC-L in some cases, which again confirms the importance of the robust GPR.

Since the main concern of this paper is the safety of RL based methods, we also list the maximum violations during online interaction in Table VI. For all the cases, the proposed S3AC achieves the lowest *VVR*, which is one or two orders smaller than other RL based methods. These results verify the enhanced safety of the proposed S3AC during online training & execution.

The computational speed is another important indicator of applicability for data-driven dispatch methods. Therefore, we test the average computation time of one step for the proposed S3AC and SAC-L, which are listed in Table VII.

TABLE VII
ONLINE COMPUTATION TIME OF ONE STEP FOR PROPOSED S3AC AND SAC-L

Test system	Average computation time (ms)		
	S3AC (safe)	S3AC (unsafe)	SAC-L
33-bus	29.8	30.7	0.735
69-bus	81.1	83.2	0.768
141-bus	415.0	420	0.792

Note: S3AC (safe) represents the computation time of proposed S3AC when the generated action $\boldsymbol{a}_{RL}$ is safe, while S3AC (unsafe) represents the computation time of proposed S3AC when $\boldsymbol{a}_{RL}$ is unsafe.

As can be observed from the results, the computation time of SAC-L only includes forward calculation of the actor network, which is the fastest in all cases. When the generated action $\boldsymbol{a}_{RL}$ is safe, the computation time for the proposed S3AC includes forward calculation of the actor network and examination of the supervisor. And when the generated action $\boldsymbol{a}_{RL}$ is unsafe, the computation time for the proposed S3AC also includes action modification of the projector. The introduced supervisor and projector increase the online computation time. In addition, from the above comparison results, the main computation time of the proposed S3AC comes from the robust GPR based supervisor, which predicts the voltage magnitude of each node and power flow of each branch in the ADN.

However, since the real-time dispatch of ADNs is usually minute-level, this computation time, which only takes milliseconds, is completely sufficient for real practice. Moreover, the ADN operator can select only critical buses and branches for safety examination, so that computation time for the supervisor can be further reduced.

V. CONCLUSION

The ADN is a safety-critical system, so the safety of the RL agent policies is a major concern to achieve the optimal dispatch in ADNs without accurate network models. One of the key novelties of the proposed S3AC is the introduction of the supervisor and projector. In the offline stage, the supervisor and projector are pre-trained with a small amount of historical data from the SCADA system. In the online stage, the supervisor and projector provide enhanced safety for every action executed on the real-world ADN. In addition, the supervisor in the proposed S3AC is formulated with a robust GPR, which is robust to outliers in measurements. Numerical studies on IEEE 33-bus, 69-bus, and 141-bus test systems have verified the safety and effectiveness of the proposed S3AC.

In future work, the application of the proposed S3AC to other safety critical problems, such as transient voltage stability, is also a possible research direction.

REFERENCES

[1] C. D'Adamo, S. Jupe, and C. Abbey, "Global survey on planning and operation of active distribution networks-update of CIGRE C6.11 working group activities," in *Proceedings of IET Conference Publications*, Prague, Czech Republic, Jun. 2009, pp. 1-4.

[2] R. Hidalgo, C. Abbey, and G. Joós, "A review of active distribution networks enabling technologies," in *Proceedings of IEEE PES General Meeting*, Minneapolis, USA, Jul. 2010, pp. 1-9.

[3] A. R. Malekpour and A. Pahwa, "Reactive power and voltage control in distribution systems with photovoltaic generation," in *Proceedings of 2012 North American Power Symposium (NAPS)*, Champaign, USA, Sept. 2012, pp. 1-6.

[4] S. Pirouzi, J. Aghaei, M. A. Latify *et al.*, "A robust optimization approach for active and reactive power management in smart distribution networks using electric vehicles," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2699-2710, Sept. 2018.

[5] X. Li, R. Ma, W. Gan *et al.*, "Optimal dispatch for battery energy storage station in distribution network considering voltage distribution improvement and peak load shifting," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 1, pp. 131-139, Jan. 2022.

[6] L. Shen, X. Dou, H. Long *et al.*, "A cloud-edge cooperative dispatching method for distribution networks considering photovoltaic generation uncertainty," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1111-1120, Sept. 2021.

[7] R. Ma, X. Li, Y. Luo *et al.*, "Multi-objective dynamic optimal power flow of wind integrated power systems considering demand response," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 4, pp. 466-473, Dec. 2019.

[8] X. Feng, S. Lin, W. Liu *et al.*, "Distributionally robust optimal dispatch of offshore wind farm cluster connected by VSC-MTDC considering wind speed correlation," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 3, pp. 1021-1035, May 2023.

[9] W. Wang, N. Yu, Y. Gao *et al.*, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Jul. 2020.

[10] Y. Zhang, X. Wang, J. Wang *et al.*, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 361-371, Jan. 2021.

[11] H. Liu and W. Wu, "Federated reinforcement learning for decentralized voltage control in distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3840-3843, Sept. 2022.

[12] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for volt-var control in power distribution networks," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3594-3604, Jul. 2021.

[13] Y. Zhou, W. Lee, R. Diao *et al.*, "Deep reinforcement learning based real-time AC optimal power flow considering uncertainties," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1098-1109, Sept. 2022.

[14] W. Liu, J. Shen, S. Zhang *et al.*, "Distributed secondary control strategy based on *Q*-learning and pinning control for droop-controlled microgrids," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1314-1325, Sept. 2022.

[15] C. Huang, H. Zhang, L. Wang *et al.*, "Mixed deep reinforcement learning considering discrete-continuous hybrid action space for smart home energy management," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 743-754, May 2022.

[16] H. Hua, X. Chen, L. Gan *et al.*, "Demand-side joint electricity and carbon trading mechanism," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 2, pp. 14-25, Nov. 2023.

[17] H. Hua, Z. Qin, N. Dong *et al.*, "Data-driven dynamical control for bottom-up energy Internet system," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 315-327, Jan. 2022.

[18] X. Shi, Y. Xu, G. Chen *et al.*, "An augmented Lagrangian-based safe reinforcement learning algorithm for carbon-oriented optimal scheduling of EV aggregators," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 795-809, Jan. 2024.

[19] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger *et al.*, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737-2752, Jul. 2019.

[20] W. Luo, W. Sun, and A. Kapoor. (2022, Jul.). Sample-efficient Safe learning for online nonlinear control with control barrier functions. [Online]. Available: https://arxiv.org/abs/2207.14419

[21] Y. Guan, Y. Ren, Q. Sun *et al*., "Integrated decision and control: toward interpretable and computationally efficient driving intelligence," *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 859-873, Feb. 2023.

[22] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2037-2047, May 2021.

[23] C. Tessler, D. J. Mankowitz, and S. Mannor. (2018, Jan.). Reward constrained policy optimization. [Online]. Available: https://arxiv. org/abs/1805.11074

[24] Y. Chow, M. Ghavamzadeh, L. Janson *et al*. (2015, Dec.). Risk-constrained reinforcement learning with percentile risk criteria. [Online]. Available: https://arxiv.org/abs/1512.01629

[25] H. Ma, C. Liu, S. Li *et al*. (2021, Nov.). Joint synthesis of safety certificate and safe control policy using constrained reinforcement learning. [Online]. Available: https://arxiv.org/abs/2111.07695

[26] J. Achiam, D. Held, A. Tamar *et al*. (2017, May). Constrained policy optimization. [Online]. Available: https://arxiv.org/abs/1705.10528

[27] T. Yang, J. Rosca, K. Narasimhan *et al*. (2020, Oct.). Projection-based constrained policy optimization. [Online]. Available: https://arxiv. org/abs/2010.03152

[28] H. Liu and W. Wu, "Online multi-agent reinforcement learning for decentralized inverter-based volt-var control," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2980-2990, Jul. 2021.

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

[30] Y. Ye, H. Wang, P. Chen *et al*., "Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial-temporal perception," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3759-3775, Sept. 2023.

[31] Y. Xia, Y. Xu, Y. Wang *et al*., "A safe policy learning-based method for decentralized and economic frequency control in isolated networked-microgrid systems," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 1982-1993, Oct. 2022.

[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge: The MIT Press, 1998.

[33] T. Haarnoja, A. Zhou, K. Hartikainen *et al*. (2018, Dec.). Soft actor-critic algorithms and applications. [Online]. Available: https://arxiv.org/abs/1812.05905

[34] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge: The MIT Press, 2005.

[35] Z. Li, L. Li, and Z. Shao. (2022, Nov.). Robust Gaussian process regression based on iterative trimming. [Online]. Available: https://paperswithcode.com/paper/robust-gaussian-process-regression-based-on

[36] M. E. Baran and F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401-1407, Apr. 1989.

[37] D. Das, "Optimal placement of capacitors in radial distribution system using a fuzzy-GA method," *International Journal of Electrical Power & Energy Systems*, vol. 30, no. 6-7, pp. 361-367, Jul. 2008.

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

[39] G. Brockman, V. Cheung, L. Pettersson *et al*. (2026, Jun.). OpenAI gym. [Online]. Available: https://arxiv.org/abs/1606.01540

[40] R. G. Cespedes, "New method for the analysis of distribution networks," *IEEE Transactions on Power Delivery*, vol. 5, no. 1, pp. 391-396, Jan. 1990.

Xu Yang received the B.S. degree from the Electrical Engineering Department, Tsinghua University, Beijing, China, in 2022, where he is currently pursuing the Ph.D. degree. His research interests include active distribution system operation and control, safe and robust reinforcement learning and its application in the energy system.

Haotian Liu received the B.S. and Ph.D. degrees from the Electrical Engineering Department, Tsinghua University, Beijing, China, in 2018 and 2023, respectively. He is currently a Postdoctor with Tsinghua University. His research interests include active distribution system operation and control, model-free control, machine learning especially reinforcement learning, and their applications in the energy system.

Wenchuan Wu received the B.S., M.S., and Ph.D. degrees from the Electrical Engineering Department, Tsinghua University, Beijing, China, in 1996, 1998, and 2023, respectively. He is currently a Full Professor with Tsinghua University. His research interests include energy management system, active distribution system operation and control, machine learning and its application in energy system.

Qi Wang received the B.S. degree from the School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree in the Electrical Engineering Department, Tsinghua University, Beijing, China. His research interests include optimization and control in hierarchical power system with integration of renewable generation.

Peng Yu received the Ph.D. degree in electrical and electronic engineering from Dalian University of Technology, Dalian, China, in 2012. He is currently with the State Grid Shandong Electric Power Company, Jinan, China. His research interests include distributed generation, microgird, and integrated energy system.

Jiawei Xing received the M. S. degree from the China-EU Institute for Clean and Renewable Energy, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2019. He is currently with the State Grid Shandong Electric Power Company, Jinan, China. His research interests include distributed generation, microgird, and integrated energy system.

Yuejiao Wang received the M. S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2015. Currently, she works at the Electric Power Science Research Institute of State Grid Shandong Electric Power Company, Jinan, China. Her research interests include distributed photovoltaic scheduling and operation management.