

Active Power Correction Strategies Based on Deep Reinforcement Learning—Part I: A Simulation-driven Solution for Robustness

Peidong Xu, *Student Member, IEEE*, Jiajun Duan, *Member, IEEE*, Jun Zhang[✉], *Senior Member, IEEE*, Yangzhou Pei, Di Shi, *Senior Member, IEEE*, Zhiwei Wang, *Senior Member, IEEE*, Xuzhu Dong, *Senior Member, IEEE*, and Yuanzhang Sun, *Senior Member, IEEE*

Abstract—Employing the novel Deep Reinforcement Learning approach, this paper addresses the active power corrective control in modern power systems. Seeking to minimize the joint effect engendered by operation cost and blackout penalty, this correction strategy focuses on evaluating the robustness and adaptability aspects of the control agent. In Part I of this paper, where robustness is the primary focus, the agent is developed to handle unexpected incidents and guide the stable operation of power grids. A Simulation-driven Graph Attention Reinforcement Learning method is proposed to perform robust active power corrective control. The aim of the graph attention networks is to determine the representation of power system states considering the topological features. Monte Carlo tree search is adopted to select the best suitable action set out of the large action space, including generator redispatch and topology control actions. Finally, driven by simulation, a guided training mechanism along with a long-short-term action deployment strategy are designed to help the agent better evaluate the action set while training and to operate more stably when deployed. The efficacy of the proposed method has been demonstrated in the “2020 Learning to Run a Power Network - Neurips Track 1” global competition and the associated cases. Part II of this paper deals with the adaptability case, where the agent is equipped to better adapt to a grid that has an increasing share of renewable energies through the years.

Index Terms—Active power corrective control, deep reinforcement learning, graph attention networks, simulation-driven.

NOMENCLATURE

B	Connected buses of power equipment in substations.
c	Constant to control the searching depth.
$C(\cdot)$	Function associated with the adjustment cost.

$C_{\text{net}}(t)$	Network loss cost.
$E_{\text{loss}}(t)$	Energy loss at time t when a blackout occurs.
h_i	Transformed features of the i th node.
K	Number of independent attention mechanisms.
$n(s_t)$	Simulation number of nodes s_t .
N_C	Number of allowed topological actions.
N_L	Number of lines.
N_S	Total simulation number.
$\mathcal{N}_i$	Neighborhood node set of the i th node.
$p(t)$	Marginal price.
$\Delta P_G, \Delta P_L$	Amount of generator redispatch and load shedding in the power system, respectively.
$P_{G \max}, P_{G \min}$	Upper and lower bounds of generator outputs.
P_{redisp}	Redispatch amount of each generator.
P, Q	Power status of the power equipment.
$R_{\text{up}}, R_{\text{down}}$	Upper and lower limits of generator ramping capabilities.
T	Length of the control period
$T_{s, \text{end}}$	Time step when the simulation ends.
$u(n(s_t))$	Bonus of each terminated simulation.
$W(s_t)$	State evaluation value of node s_t .
W^k	k th weight matrix.
$X_{\text{line}}, X_{\text{bus}}$	Number of line-switching actions, bus bar switching actions.
α_{ij}^k	Normalized attention coefficients computed by the k th attention mechanism.
α, β	Penalty factors of overload and heavy load, respectively.
π	Shared attentional mechanism.
ρ	Load ratio of each powerline.
σ	Non-linear activation function.
Ω	Predefined threshold of load ratio.

Manuscript received December 30, 2020; revised February 18, 2021; accepted April 14, 2021. Date of online publication September 10, 2021; date of current version November 8, 2021. The work is supported by the National Key R&D Program of China under Grant 2018AAA0101504 and the Science and technology project of SGCC (State Grid Corporation of China): fundamental theory of human-in-the-loop hybrid-augmented intelligence for power grid dispatch and control.

P. D. Xu, J. Zhang (corresponding author, e-mail: jun.zhang.ee@whu.edu.cn); ORCID: 0000-0001-6908-2671), Y. Z. Pei, X. Z. Dong, and Y. Z. Sun are with the School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China.

J. J. Duan, D. Shi, and Z. W. Wang are with GEIRI North America, San Jose, CA 95134, USA.

DOI: 10.17775/CSEEJPES.2020.07090

I. INTRODUCTION

THE extensive implementation of ultra-high voltage transmission technology and renewable energy integration often necessitates the power grids to operate under large-scale and long-distance power exchange conditions, making it

highly demanding to uphold the stable operation of modern power systems. In the event of power line overloads, it becomes essential to carry out elimination or alleviation of the overloads, or else subsequent cascading failure may be triggered. Hence, keen attention must be given to the analysis of fast and efficient active power corrective control techniques, to maintain safe and stable operation of the power networks.

The typical corrective control methods emphasize primarily on the generator redispatch or load shedding actions, where the node injections are optimized to adjust line power flow [1]. Owing to the progress of modern power systems, numerous additional control measures can be introduced to fix the line overloads. In [2], Flexible AC Transmission System (FACTS) devices are utilized to perform day-ahead corrective control for Security-Constrained Unit Commitment (SCUC) problems. Switching of the multi-terminal direct current (MTDC) power and series capacitors (SC) is utilized to relieve line overloads [3], [4]. Depicting promising flexibility, topology controls, such as transmission switching (TS) and substation configurations, are considered effective power flow control technologies and are known to yield excellent cost reduction in economic dispatch and transmission congestion management [5]–[9]. Furthermore, the application of corrective TS aids in preserving the ramping capability of generators, thereby making the system more robust [10]. In [11]–[13], it is demonstrated that topology actions tend to be reliable for corrective overload control. Accordingly, line and bus bar switching can be adopted as feasible corrective control measures coupled with conventional tactics.

Currently, the commonly used corrective control methods can be categorized into sensitivity-based and optimization-based methods. The former category takes into account sensitivity factors to quantify the effects produced by certain measures on the overloaded lines and opts for high-sensitivity measures to handle the overloads; even though it appears to be computation-efficient and readily executable, it may suffer from low accuracy attributable to the adoption of the direct current (DC) power flow method [14]. As a result, sensitivity-based methods are often combined with other methods to enhance their accuracy. Likewise, in [15], sensitivity factors are introduced to select the most effective load shed buses for further optimization. For the optimization-based methods, corrective control measures are essentially selected by interpreting optimization models. By formulating the optimal corrective line switching problem in the form of a constraint satisfaction problem, constraint programming and tree search methods are adopted to find a solution to the given problem [16]. In [17], the robust corrective switching problem is expressed as a mixed integer programming (MIP) problem, and hence, a robust optimization approach is adopted to deal with the accompanying uncertainties. Moreover, certain empirical methods are also brought into play to obtain optimal solutions [18], [19]. Although the optimization-based methods can readily adapt to the intricate constraints of power systems, they may be computationally expensive. In [20], an approach based on Radial Basis Function Neural network (RBFN) is put forward to achieve successful mitigation of line overload.

In recent times, the rapid development and extensive ap-

plication of reinforcement learning (RL) have transformed it into a highly viable and much-preferred solution opted for multiple decision and control problems across different time scales and electrical system states [21]–[23]. Therefore, RL can be used to perform time-series corrective control for the stable operation of intricate modern power systems. In our previous work, the deep reinforcement learning (DRL) method was utilized to perform autonomous voltage and line flow control [24]–[26].

Hence, based on our previous work, a simulation-driven reinforcement learning method in combination with the graph neural networks is presented in this paper for suitably realizing active power corrective control. Topology actions and generator redispatch aspects are combined to devise comprehensive corrective measures for the DRL agent. To address the two challenges of complex observation space and action space in our problem, Graph Attention Networks (GATs) and the Monte Carlo tree search (MCTS) method are adopted to resolve the above-mentioned issues, respectively. Several sub-agents are determined to perform analogous training and collaborative deployment in accordance with various action sub-spaces. A reward-guided deep exploration mechanism is developed for the sub-agents to grasp correction actions effectively. The long-short-term action deployment strategy is proposed to select output action from each sub-agent to balance the immediate control effect and long-term benefit. The efficacy of the proposed method is ultimately verified through a global power system AI competition and its associated cases.

The remainder of this paper is organized as follows: Section II presents the active power correction problem and formulates it as a Markov decision process (MDP). Section III illustrates our approach to enhancing the efficacy and performance of the control agent. Case studies are given in Section IV to demonstrate the procedure of the proposed method. Section V summarizes our work and represents a picture of our future directions.

II. PROBLEM FORMULATION

A. Objective and Constraints

The objective of active power corrective control is generally described as:

$$\min C(|\Delta P_G|, \Delta P_L) \quad (1)$$

where $C(\cdot)$ denotes the function associated with the adjustment cost, and ΔP_G and ΔP_L represent the amount of generator redispatch and load shedding in the power system, respectively. It is noteworthy that the topological changes are presumed to be cost-free here.

In addition to the traditional power system constraints, the number of topological control actions and redispatch amounts of generators must also be curtailed to ensure minimum disturbance to the power system.

$$X_{\text{line}} + X_{\text{bus}} \leq N_C \quad (2)$$

$$\begin{cases} |\Delta P_G| \leq \min(P_{G_{\max}} - P_G, R_{\text{up}}) \\ |\Delta P_G| \leq \min(P_G - P_{G_{\min}}, R_{\text{down}}) \end{cases} \quad (3)$$

where X_{line} , X_{bus} , N_C symbolize the number of line-switching actions, bus bar switching actions, and the allowed topological actions, respectively. $P_{G \max}$, $P_{G \min}$, R_{up} , R_{down} denote the upper and lower bounds of the generator outputs, as well as the upper and lower limits of the generator ramping capabilities.

As the active power correction focuses on eradicating or at least alleviating the line overloads in an effort to avert cascading failures and maintain the operation stability of the power systems, in this paper, the problem is extended to a continuous real-time power system control problem. Based on the carefully planned dispatch schedule, corrective actions are executed when the normal overloads or post-contingency overloads occur; “do-nothing” action is carried out when no line violates the constraints. The objective then gets transformed into maintaining the system operation while minimizing the overall cost during the entire period as:

$$\min \sum_{t=0}^T [C(|\Delta P_G(t)|, \Delta P_L(t), t) + C_{\text{net}}(t) + E_{\text{loss}}(t) \cdot p(t)] \quad (4)$$

where T denotes the length of the control period, $C_{\text{net}}(t)$ represents the network loss cost, which also reflects the economic influence of corrective actions. $E_{\text{loss}}(t)$ symbolizes the energy loss at time t when a blackout occurs and $p(t)$ represents the marginal price of the generators’ productions. By combining operation cost and blackout penalty, the optimal control strategy becomes proficient enough to maintain the stability of the power grid using less costly corrective actions.

Finally, apart from the above constraints, reaction time, i.e., time to react to an overloaded line before it gets disconnected by relay protectors, and recovery time, i.e., the time required to reactivate the power equipment after its activation, are introduced to strengthen the feasibility of the optimized strategy in time-series control.

B. Problem Formulated as MDP

As the objective function shown in (4) requires a time-series control manner, the active power correction can be modeled as MDP in the form of a 5-element tuple $M = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\}$. Among these elements, γ denotes a discounting factor and $\mathcal{P}$ represents the transition probability matrix. The other elements in our problem are elaborated as follows:

State space $\mathcal{S}$: The agent state $s_t \in \mathcal{S}$ represents the observation attained by the power system. As we concentrate on the alleviation of line overloads by node injections and topological changes, the features of generators, loads, and both ends of the transmission lines must be considered. This state contains active power status, reactive power status, bus connection of the power equipment, redispatch status of the generators, as well as load ratio of each line, i.e.,

$$s_t = (P, Q, B, P_{\text{redispatch}}, \rho) \quad (5)$$

where P and Q denote the power status of the power equipment, including power outputs of generators, consumption value of loads, and power flow at both ends of the lines. B signifies the connected buses of power equipment in substations. $P_{\text{redispatch}}$ represents the redispatch amount of each generator. ρ

denotes the load ratio of each powerline, i.e., the current flow divided by the thermal limit of each powerline.

Action space $\mathcal{A}$: The action space comprises of generator redispatch, line switching, bus bar switching, and do-nothing actions. The line switching changes the state of the power grids by reconnecting/disconnecting a powerline. The bus bar switching entails the bus selection of elements connected to the bus bar, where each element can be switched onto either of the buses of the bus bar, to control the power system better.

Reward function $\mathcal{R}$: As our principal objective remains to maintain the operational stability of the power system under the likelihoods of load and renewable energy generation fluctuation, maintenance, as well as multiple other contingencies, hence the residual capacity of the grid must be optimized to enhance the grid flexibility in the case of unexpected events. Furthermore, since loaded and overloaded lines may become possible chokepoints to power transmission, their existence should be considered as a penalty on the flexibility of the system. Thus, the performance at time t can be formulated as [27]:

$$o_t = \sum_{i=1}^{N_L} [\max(0, (1 - \rho_i^2)) - \alpha \cdot \max(0, \rho_i - 1) - \beta \cdot \max(0, \rho_i - 0.9)] \quad (6)$$

where N_L denotes the number of lines, and α and β represent the penalty factors of overload and heavy load, respectively.

Concerning this problem, the system tends to act controlled in a time-series manner, and the power system is expected to survive longer. Hence, if the DRL agent is rewarded both by the past as well as the current performance, the agent will be motivated to select more efficient actions to evade any harm to the system stability maintained from the past to the present time step. Thus, the immediate reward r_t can be defined as:

$$r_t = \begin{cases} \lambda & \text{if system fails} \\ \sum_{i=0}^t o_t & \text{otherwise} \end{cases} \quad (7)$$

where λ is a negative constant.

III. SIMULATION-DRIVEN GRAPH ATTENTION REINFORCEMENT LEARNING METHOD

Of late, reinforcement learning has emerged as a great solution for tackling complex decision-making problems owing to its notable capacity for the same. While performing corrective control for large power systems prone to undergoing unexpected incidents by generator redispatch and topological adjustments, there are several crucial issues to be addressed: 1) Topological changes due to random events; 2) high dimensions of topological action space; 3) high stability requirement.

To resolve the above-mentioned issues, a simulation-driven graph attention reinforcement learning (SGA-RL) method is developed in this paper, and its architecture is shown in Fig. 1. In this architecture, GATs are introduced to understand the representation of the power grid state and strengthen the agent’s generalization ability for aptly sensing unpredictable topological changes. Subsequently, the MCTS method is employed to limit the action space. Finally, a reward-guided training

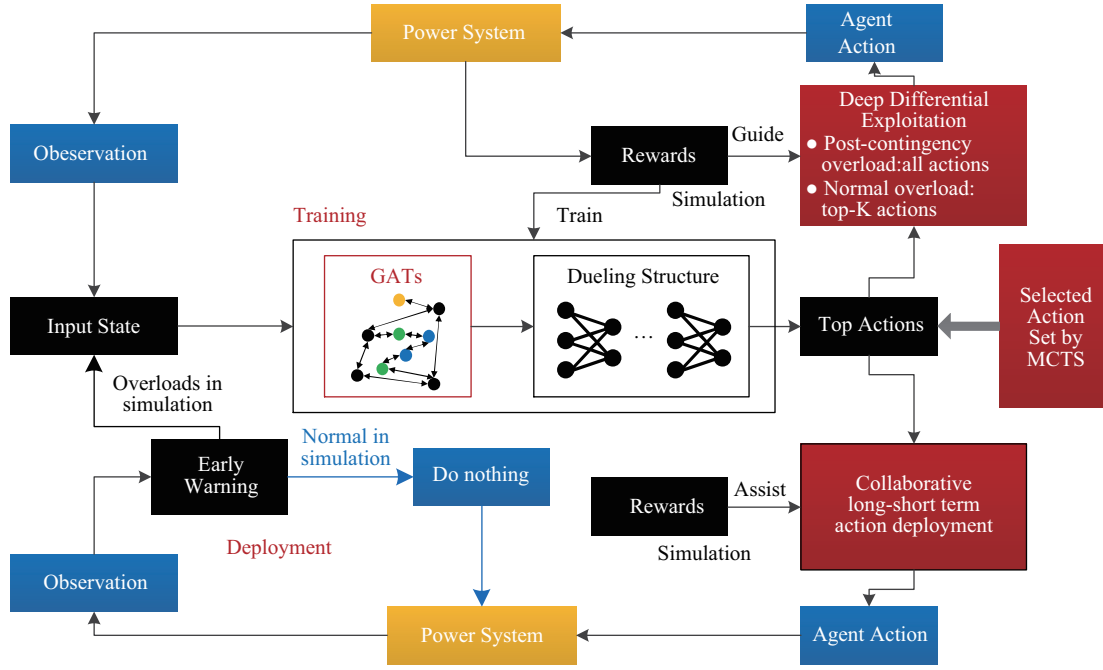


Fig. 1. The architecture of the proposed method.

mechanism and a long-short-term action deployment strategy are formulated for accomplishing effective corrective control strategy learning and its stable application, respectively. To prevent the power system from defying the constraints, an early warning mechanism is installed as well [25], i.e., when the maximum load ratio in the simulation exceeds the predefined threshold Ω , e.g., 0.98, the agent is triggered to undertake control decisions, to protect the system in a preventive and corrective manner. Double-dueling deep Q-network is also adopted as the basic DRL framework.

A. Graph Attention Networks (GATs)

Graph neural networks (GNNs) are utilized to extract features from the power grid states. The state is transformed into a graph structure data, as in [26]. The adjacency matrix $\mathbf{A}$ can be formed based on the connections between the power equipment, i.e., ends of lines, generators, and loads. The feature matrix $\mathbf{X}$ is formed by combining the common and the unique features of various equipment. An example is shown in Fig. 2 and (8).

$$\left\{ \begin{array}{l} \mathbf{A} = \begin{array}{l} \text{Origin} \\ \text{Extremity} \\ \text{Load} \\ \text{Generator} \end{array} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \\ \mathbf{X} = \begin{array}{l} \text{Origin} \\ \text{Extremity} \\ \text{Load} \\ \text{Generator} \end{array} \begin{bmatrix} P_{\text{OR}} & Q_{\text{OR}} & B_{\text{OR}} & 0 & \rho \\ P_{\text{EX}} & Q_{\text{EX}} & B_{\text{EX}} & 0 & \rho \\ P_{\text{L}} & Q_{\text{L}} & B_{\text{L}} & 0 & 0 \\ P_{\text{G}} & Q_{\text{G}} & B_{\text{G}} & P_{\text{resisp}} & 0 \end{bmatrix} \end{array} \right. \quad (8)$$

where P_{OR} and Q_{OR} denote the active power flow and the reactive power flow at the origin of the transmission line, P_{EX} and Q_{EX} represent the active power flow and the reactive

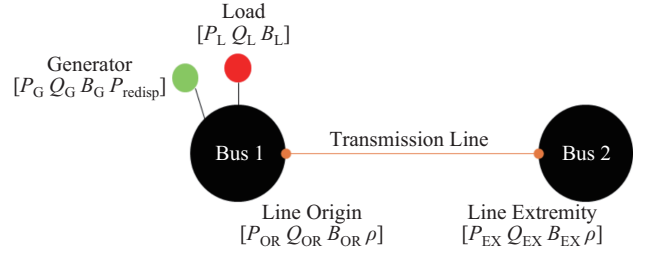


Fig. 2. Example for the establishment of the adjacency matrix and feature matrix.

power flow at the extremity of the transmission line, P_{L} and Q_{L} represent the active consumption value and the reactive consumption value of the load, and P_{G} and Q_{G} stand for the active power output and the reactive power output of the generator. B_{OR} , B_{EX} , B_{L} , and B_{G} symbolize the connected buses of the above power equipment at the substation. P_{resisp} is the redispatch amount of generator, and ρ denotes the load ratio of the transmission line.

In the problem considered here, when the grid undergoes maintenance or “ $N - 1$ ” events occur, the resultant topologies end up being quite different from the normal ones. Therefore, the DRL agent must possess satisfactory generalization ability to take decisions under unanticipated topologies caused by random events and to perform better in various operation scenarios.

Thus, GATs are utilized to carry out representation learning for power system status. Unlike graph convolutional networks (GCNs), which depend on the Laplacian eigenbasis, in GATs, the hidden representations of each node in the graph can be computed following a masked self-attention strategy, where the transformed features of the target node are accumulated solely based on the information derived from its neighboring

nodes. Thus, the models employing GATs can be generalized to analyze different structures. The graph attention layer with multi-head attention for transformed features of the i th node h_i is defined in (9) as [28]:

$$h'_i = \sigma \left(\frac{1}{K} \sum_{k=1}^K \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k h_j \right) \quad (9)$$

where σ represents the non-linear activation function, $\mathcal{N}_i$ denotes the neighborhood node set of the i th node, K represents the number of independent attention mechanisms, α_{ij}^k symbolizes the normalized attention coefficients computed by the k th attention mechanism [29], and $\mathbf{W}^k$ is the k th weight matrix.

Specifically, in the k th attention mechanism, the weight matrix $\mathbf{W}^k$ is applied to each node to execute self-attention. The importance of neighboring j th node's features for i th node can be calculated based on the shared attentional mechanism: $\pi(\mathbf{W}^k h_i, \mathbf{W}^k h_j)$. By adopting the softmax function, α_{ij}^k can be obtained. As we combine the normalized attention coefficients of K attention mechanisms in (9), the multi-head attention is realized and the aggregated features of i th node are produced.

B. Monte Carlo Tree Search (MCTS)

In large-scale power systems, a large number of available corrective actions exist due to the discreteness and combination of the topological adjustments. For instance, if 2 buses are on a bus bar, and n elements are connected to the substation, there will be 2^{n-1} possible actions. The huge action space makes the double-dueling DQN agent challenging or even impossible to train.

To reduce the dimension of the action space, the MCTS method is adopted to select the effective actions by sampling and evaluating the entire action space. Typical scenarios are chosen to form the searching set, and in each scenario, the “do-nothing” or “auto-reclosing” action is executed unless overload occurs. The state with line overload is considered to be the root node, and then the selection-expansion-simulation-backpropagation process can be performed to update the tree [30].

To obtain effective corrective actions, the leaf nodes of the tree represent the next overload states after the application of feasible actions in the previous overload states, while the edges represent the corrective actions a_t . Each node s_t of the tree stores its state evaluation value as $W(s_t)$ and simulation number as $n(s_t)$. $W(s_t)$ is initially set to 0 and updated by the bonus $u(n(s_t))$ of each terminated simulation. The bonus $u(n(s_t))$ evaluates the equivalent effect of corrective action a_t in the corresponding simulation, which can be defined in (10) as:

$$\begin{cases} u(n(s_t)) = \frac{T_{s,\text{end}} - t}{T - t} \\ W(s_t)_{n(s_t)} = W(s_t)_{n(s_t)-1} + u(n(s_t)) \end{cases} \quad (10)$$

where $T_{s,\text{end}}$ denotes the time step when the simulation ends.

At the early stage of the MCTS process, the enumeration of available generators, switchable branches, and buses is performed for the current overload state, to form the effective

action set. After the enumeration is completed, in the subsequent MCTS process, action selection follows the UCT (upper confidence bound applied to trees) to reach the node with the best performance calculated by (11). The overall structure of MCTS in this paper is illustrated in Fig. 3.

$$\text{score} = \frac{W(s_t)}{n(s_t)} + c \sqrt{\frac{\ln N_S}{n(s_t)}} \quad (11)$$

where N_S denotes the total simulation number, and c is a constant to control the searching depth.

C. Reward-guided Training Mechanism

After viable actions are identified, the action space still possesses high dimensions for bulk power systems or complex substation configurations. To further enhance the training efficiency of the DRL agent, the action space is divided into F sub-spaces, and F sub-agents are established to perform training under the same mechanism. In addition, at each time step, the prior auto-reclosing check is carried out before the execution of the sub-agent's action. If there exist disconnected lines available to be reconnected, the auto-reclosing action is performed to reduce the abnormality of topology caused by the unanticipated events and alleviate the control difficulty for the sub-agent.

Since line overloads often cause cascading failures, the corrective actions need to be effective. Thus, in overload states, δ actions with the highest Q-values are selected and examined by the simulation system in descending order. The feasible action with the highest simulated reward is chosen as the final decision. When the grid is in a post-contingency state, the loaded or the overloaded lines can prove to be major safety hazards for the grid stability, hence here, δ equals to the number of entire sub-agents' action space, i.e., enumerating all the actions to find an action with the best reward. The reward-guided differential exploration strategy facilitates the sub-agent in examining the action set more precisely, especially in a post-contingency state, to execute valid actions when deployed.

In most operational states, “do-nothing” is carried out by the DRL model owing to its pre-dispatching schedule. While applying experience replay, many “do-nothing” transitions are stored in a replay buffer, leading to the low sample frequency of other useful transitions to perform experience replay, which impedes the proper evaluation of the corrective control actions. In this paper, a balanced replay buffer is designed: The “do-nothing” transitions are stored in the buffer with a minimal probability to maintain the proportion of transitions related to redispatch and topological control actions in the buffer.

The algorithm for training simulation-driven graph DQN agents is given in Algorithm 1.

D. Long-Short Term Action Deployment Strategy

In this paper, we expect an optimal active power correction strategy to eliminate the overloads or at last reduce them effectively and perform a continuous control for minimizing the overall costs considering blackout penalties. Therefore, both the short-term control effects and long-term benefits of the output actions during the DRL model's implementation should be considered.

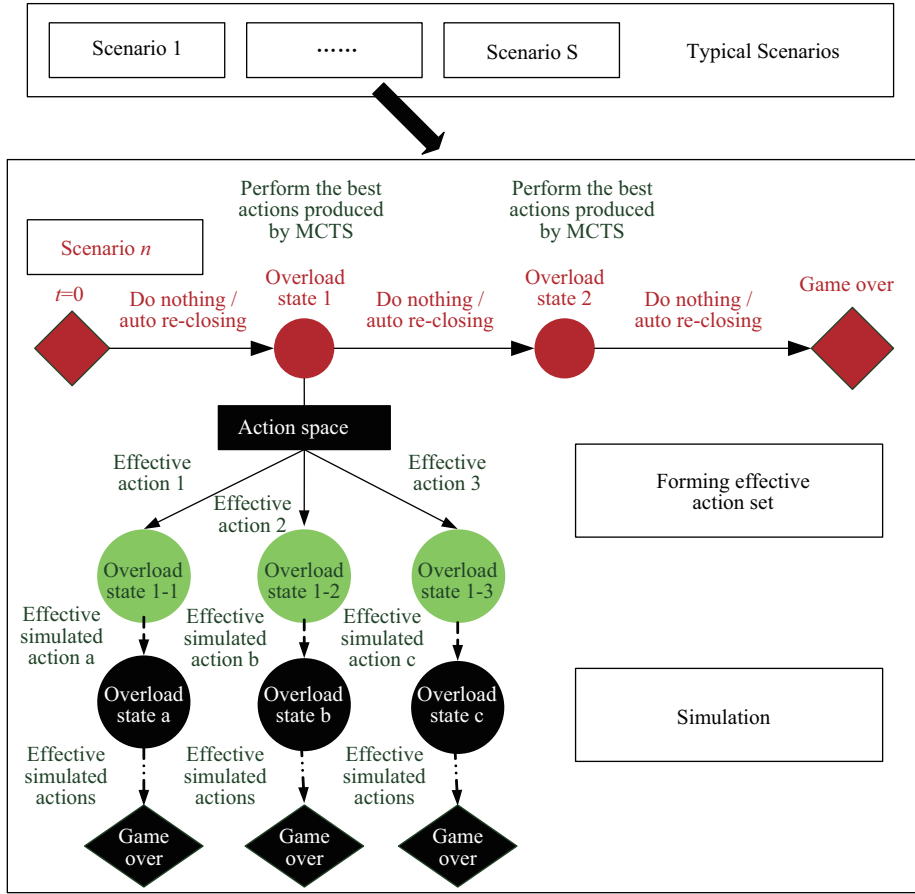


Fig. 3. Monte Carlo tree search in our problem.

Based on the collaboration of all the trained sub-agents, a long-short-term action deployment strategy is devised. At overload state s_t , in each sub-agent, top- k actions with the largest Q-values are selected and verified by the simulation system. Noticeably, k varies in different types of overloads.

In the 1st sub-agent, the verified control action with the highest Q-value is chosen as the base action to serve as the action baseline with high long-term benefit, as defined in (12):

$$a_t^{\text{base}} = \arg \max_a Q_1(s_t, a_{\text{simu}}) \quad (12)$$

The base action is then updated by other sub-agents according to the simulated reward to promote the better short-term

control effect of the final action. For one sub-agent, its verified top actions are compared to the base action by simulated rewards. When a top action overcomes the base action by the simulated reward, then the comparison terminates, and the action is set as the new base action for the latter sub-agent to be compared and updated. The final base action is selected as the output action. In this manner, the output control action achieves a good immediate control effect while maintaining long-term benefit.

IV. CASE STUDIES

A. Experimental Setup

The open-source platform Grid2Op [31] is employed as the environment for our agents. A modified 36-bus system selected from the IEEE 118-bus system serves as the grid under investigation, as shown in Fig. 5. The target grid consists of 59 transmission lines, 22 generators, and 37 loads. The data set of the target grid includes 576 scenarios spanning over 12 months of 48 years. Each scenario contains data for 28 continuous days in a 5-minute resolution. Attacks on random transmission lines tend to occur every day at different times, i.e., an “ $N - 1$ ” event occurs such that the corresponding lines suffer outage for an uncertain duration. Meanwhile, the maintenance of the lines also takes place. In the platform, 24 one-week validation scenarios that have not been previously experienced are provided to evaluate the trained agents. These

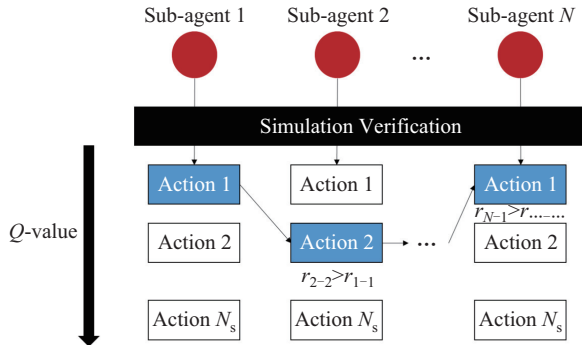


Fig. 4. Long-short-term action deployment strategy.

Algorithm 1: Simulation-driven Graph DQN Training Method

```

1 Initialize weights of Graph DQN  $\theta$ 
2 Initialize Memory Buffer  $\Delta$ 
3 for  $episode = 1$  to  $I$  do
4   Reset the environment
5   for  $t = 1$  to  $T$  do
6     Collect and transform the observation  $s_t$ 
7     Perform do-nothing action  $a^0$  in the simulation
        system, check the danger flag (True for
        overload or system failure, False for normal
        state)
8     if the danger flag is True then
9       Select  $\delta$  top actions according to  $Q(\cdot|s_t)$  of
        the agent:
        
$$\delta = \begin{cases} |\mathcal{A}| & \text{if post-contingency state} \\ V & \text{otherwise} \end{cases}$$

10      Validate the actions in the simulation
        system and choose the feasible action with
        the best reward as the output action  $a_t$ 
11    else
12      select do-nothing action  $a^0$  as the output
        action  $a_t$ 
13    end if
14    Execute action  $a_t$  in the environment and
        obtain next state  $s_{t+1}$ , reward  $r_t$ , and  $d_t$ 
15    Store the transition  $(s_t, a_t, r_t, s_{t+1}, d_t)$  in  $\Delta$ ,
        if  $a_t = a^0$ , save with a small probability  $p_s$ , if
         $d_t$  is True, save multiple times
16    Sample a batch of transitions from  $\Delta$  using
        importance sampling
17    Calculating the target Q-values:
        
$$y_i = \begin{cases} r_i & \text{if } d_t \text{ is True} \\ r_i + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) & \text{otherwise} \end{cases}$$

18    Update Q-network by losses:
         $L_i(\theta) = (y_i - Q(s_t, a_t; \theta))^2$ 
19    Hard copy main network weights  $\theta$  to the
        target network regularly
20    Update state  $s_t = s_{t+1}$ 
21  end for
22 end for

```

scenarios are selected from each month of the year with special considerations [32].

The number of topological control actions in our research is restricted to 1, i.e., $X_{\text{line}} + X_{\text{bus}} \leq 1$, to evade any unnecessary dramatic disturbances to the system. Even with this constraint, there are more than 60,000 available control actions in the 36-bus system, where most of them are bus bar switching actions. The reaction time and recovery time mentioned in Section II are both set to 3 time steps, i.e., 15 minutes.

The agents are trained on a Linux server with 4 GPUs, with each GPU having 11 GB memory.

B. Performance of the Proposed Method

The proposed simulation-driven graph reinforcement learn-

ing method is adopted to train the DRL agents for corrective control. After applying the MCTS method, around 1500 eligible actions are selected. The entire action space is divided into 4 parts, with each part studied by a corresponding sub-agent. Each sub-agent is a double-dueling DQN sub-agent consisting of 2 graph attention layers—1 dense layer and 1 dueling structure. The structure of the deep network is as shown in Fig. 6, and 4 attention mechanisms are utilized in each graph attention layer. Each sub-agent is trained for 1000 episodes and collaborates in the deployment process. In deployment, 100 top actions (when normal overloads occur) or all top actions (when post-contingency overloads occur) of each sub-agent are chosen to be validated. The “do-nothing” transitions are stored with a probability equal to 0.05.

During the training, the system operating duration of each episode, i.e., the number of time steps during which the power system remains stable before the blackout, is also recorded. The cumulative rewards curve and system operation duration curve of a sub-agent during training are shown in Fig. 7. The cumulative rewards and system operating duration are averaged and leveled out.

Considering the diversity of the training scenarios, e.g., different generation and load status, stochastic “ $N-1$ ” events, and reward-guided exploration process of sub-agents, fluctuations are bound to occur during the cumulative rewards and system operating duration. Nonetheless, according to the smoothened curves in Fig. 7, the sub-agent tends to enhance its performance during the first 200 episodes and maintains the performance for the rest of the training period.

Once the training process is completed for all 4 sub-agents, collaborative decision making is achieved by combining the outputs from these sub-agents with a predefined threshold value of $\Omega = 0.98$. A local validation scenario is selected to evaluate the graph DQN agents, while a “do-nothing” agent is also analyzed to serve as a baseline. The overall system operation status and agents’ decisions are as shown in Fig. 8.

During the one-week scenario of 2016 time steps, 7 “ $N-1$ ” events occur; the starting time steps are 96, 310, 791, 1109, 1317, 1501, and 1945, respectively. The graph DQN model survives the entire scenario, while the “do-nothing” agent only manages to run the system for 105 time steps, i.e., it ceases to work during the first “ $N-1$ ” event. Particularly, under the control of graph DQN model, 5 overloads occur during the operation, 2 of which are triggered by the first and last “ $N-1$ ” events, respectively, while the rest are normal overloads, and all the overloads are eliminated within the given reaction time.

We elaborate on the elimination of the first post-contingency overload to demonstrate the control process of the trained DRL model. As shown in Fig. 9, at time step 96, the transmission line 56 (bus26-bus30) is out of service due to a random fault, and the load ratio of line 41 (bus26-bus28) increases from 0.496 to 1.101, leading to the overload. Our model performs a bus bar switching action on bus 28: For generator 16, the extremity of line 42 and origin of line 57 are connected on bus 28-1, and the extremity of line 41 and origin of line 44 are assigned to bus 28-2. After deploying the action, the load ratio drops to 0.865, and the overload is eliminated without recurrence. Notably, from Fig. 8(b), we can observe that the

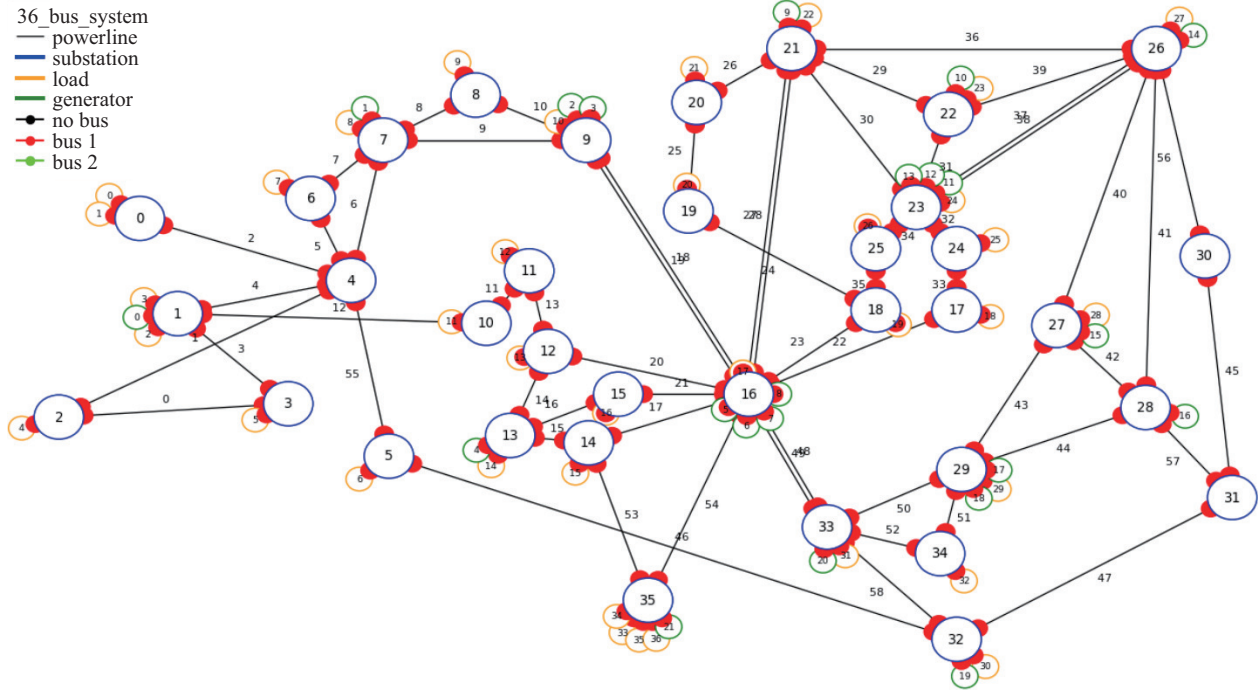


Fig. 5. The target 36-bus system.

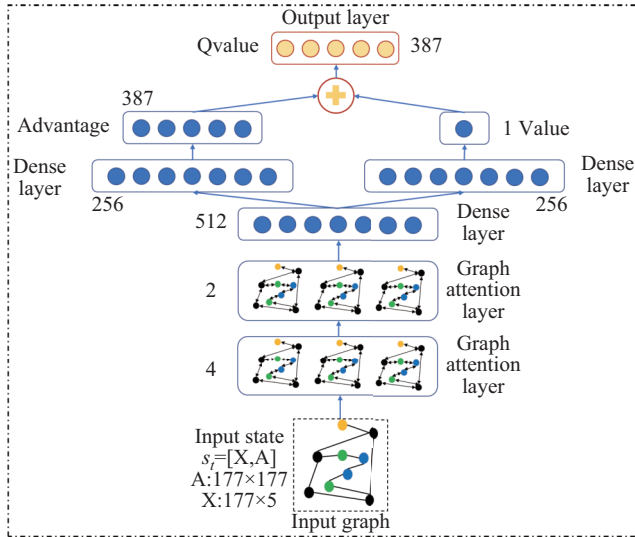


Fig. 6. The architecture of graph DQN.

overload does not disappear and can lead to system failure later if no control measure is taken. Thus, the corrective action undertaken by our model is absolutely essential and effective.

To further analyze the proposed DRL model, the 24 validation scenarios mentioned in Section IV.A are introduced. The performance is the summation of the overall cost calculated by (4) of each episode. To make it more comprehensible, the overall cost is presented as follows: Score 0 represents the performance of the “do-nothing” agent, while Score 100 stands for the performance of the best possible agent who handles all the scenarios successfully with topological actions and minimal network loss. The trained DRL model is submitted to the Robustness Track of the L2RPN competition, and another

24 unseen scenarios are provided by the competition host to test our model; the results are as shown in Table I.

TABLE I
PERFORMANCE OF PROPOSED MODEL ON VALIDATION/TEST SCENARIOS

Index	Validation scenarios	Test scenarios
Average operating steps	1255	1269
Completed episodes	8	10
Scores	42.95	43.16

As shown in Table I, the trained model can maintain the system for 4-day long stable operation on average while facing unanticipated incidents. Besides, the performance of validation and test scenarios barely show any difference, with the latter even slightly outperforming the former, highlighting the generalization ability of our model. The average decision time for each time step of our model is around 35 ms. The results exhibit the potential of the proposed method in alleviating overloads for complex modern power systems.

C. Efficacy of Graph Attention Networks (GATs)

To inspect the feasibility of adopting GATs, a GCN-based model is utilized for comparison. The two models differ from each other in the utilization of the GNNs. Two graph convolution layers are adopted in the GCN-based model for comparison. After adopting different predefined thresholds, the best GCN model is selected and compared with the GAT model on validation scenarios, which is shown in Fig. 10 and Table II.

From Fig. 10 and Table II, we can observe that both the models performed well on the validation set, and the GAT model is conducive to supporting the power system’s constant operation, thereby achieving better benefits in terms of stability and economy. Hence, it is recommended to introduce GATs

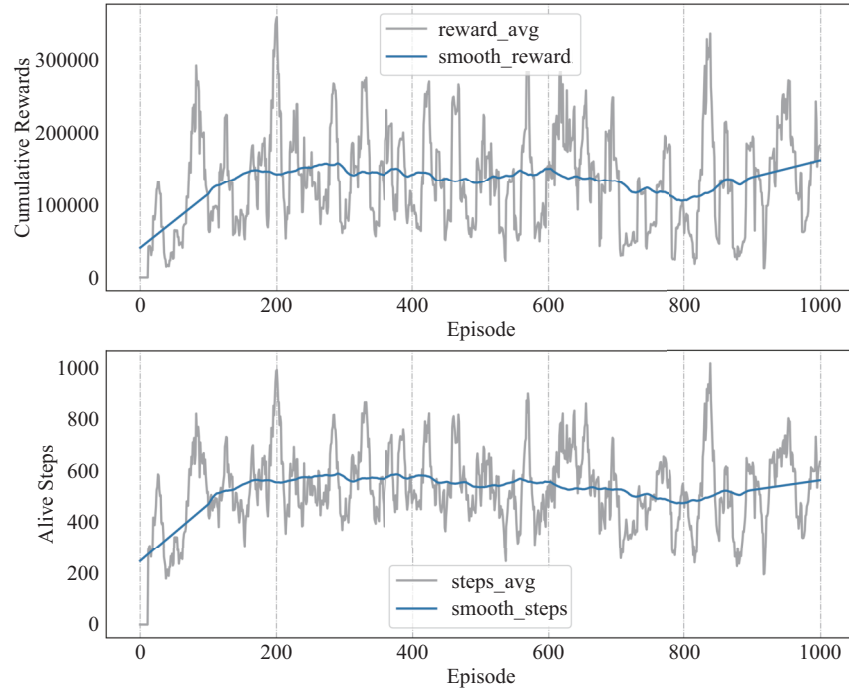
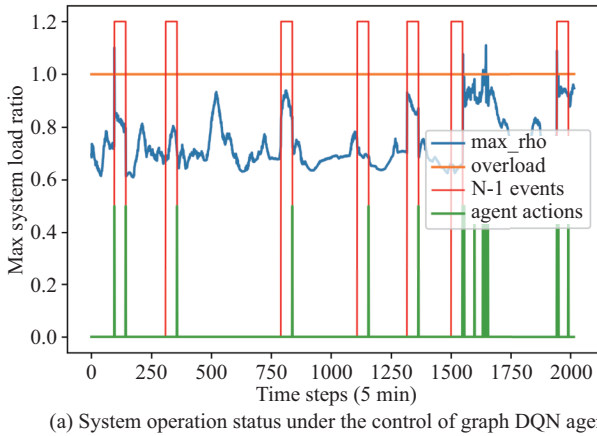
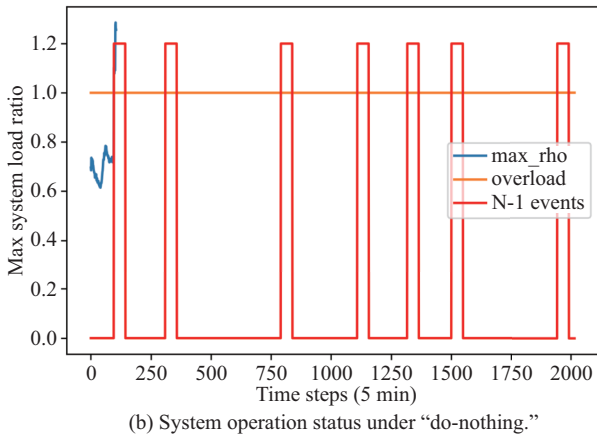


Fig. 7. The training process of the graph DQN agent.



(a) System operation status under the control of graph DQN agent.



(b) System operation status under "do-nothing."

Fig. 8. System operation status comparison.

in the RL model in case the power grid must undergo large topological changes.

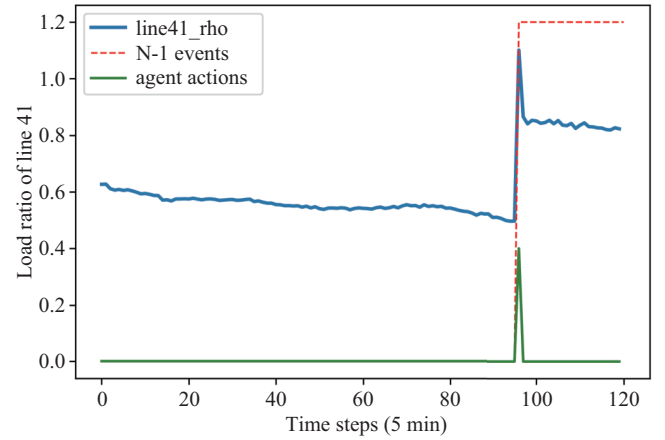


Fig. 9. Status of line 41 during the first "N - 1" event.

TABLE II
PERFORMANCE COMPARISON BETWEEN DIFFERENT MODELS

Index	GAT model	GCN model
Average operating steps	1255	1208
Completed episodes	8	7
Scores	42.95	40.48

D. Efficacy of Long-Short-Term Action Deployment Strategy

Considering the trained sub-agents, a fully reward-guided action deployment strategy and an enumeration strategy are applied to compare the performance of our long-short-term action strategy. In the fully reward-guided strategy, 100 plus top actions are simulated for each sub-agent to find the action with the best reward, to eliminate normal and post-contingency overload, respectively. In the enumeration strategy, all the sub-agents' actions are enumerated to figure out the best action

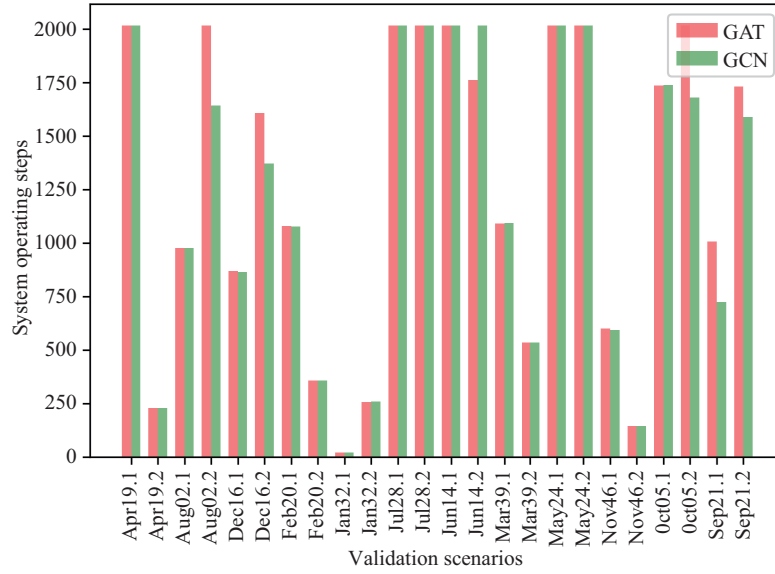


Fig. 10. System operation duration under different models.

available with the largest immediate reward.

Firstly, the two strategies are applied in the local validation scenario utilized by the long-short-term strategy in Section IV.B. The system operation status under different strategies is shown in Fig. 11 and Table III.

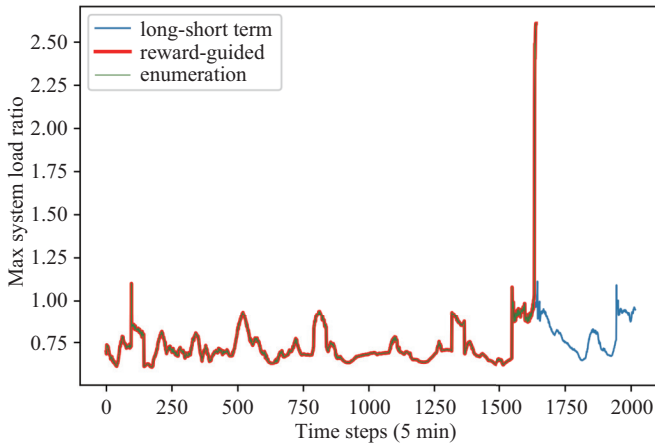


Fig. 11. System operating status under different strategies.

TABLE III
OPERATING STATUS UNDER DIFFERENT STRATEGIES

Strategy	Operating steps	Action numbers (From 1–1640)
Long-short-term	2016	15
Reward-guided	1640	21
Enumeration	1640	21

From the results in Fig. 11 and Table III, we can observe that the reward-guided action deployment method and enumeration strategy are unable to handle the overload in the scenario as properly as our proposed long-short-term method, causing severe overload and eventually leading to the blackout at time step 1640. Moreover, fewer control actions are executed under our strategy during the same period, illustrating the ability of

our strategy to be stable and effectively control the power grid by considering long-term benefits and short-term rewards.

As a further comparison, the three strategies are also evaluated under the same validation scenarios. The results are depicted in Table IV.

TABLE IV
PERFORMANCE COMPARISON BETWEEN DIFFERENT STRATEGIES

Index	Long-short-term	Reward-guided	Enumeration
Average operating steps	1255	1039	960
Completed episodes	8	6	5
Scores	42.95	32.38	26.90

As illustrated in Table IV, based on the selected action set, long-short-term action deployment strategy can dominate the strategies that emphasize instant control effects in the long-term active power control while maintaining the feasibility of corrective actions.

E. Performance Comparison of Different Predefined Threshold

In this paper, considering the intricacy of the problem, the predefined threshold Ω is utilized to provide the agents with preventive control ability. Corresponding models are established with thresholds Ω ranging from 0.95 to 1.00 to analyze their relevant performance. The results from validation scenarios are shown in Table V.

TABLE V
PERFORMANCE COMPARISON BETWEEN DIFFERENT THRESHOLDS

Ω	Average operating steps	Completed episodes	Scores
$\Omega = 0.95$	1169	8	37.16
$\Omega = 0.96$	1059	7	32.33
$\Omega = 0.97$	1163	6	35.54
$\Omega = 0.98$	1255	8	42.95
$\Omega = 0.99$	1122	7	34.64
$\Omega = 1.00$	1204	9	39.02

As shown in Table V, the different thresholds that trigger the agents can have a relatively strong influence on their per-

formance. If the control strategy is too preventive, redundant actions may disturb the power system even before the events occur. However, a high threshold may prevent the agents from taking due action in time, owing to the strict constraints. Hence, a suitable threshold must be selected, keeping in mind the system characteristics and constraints.

V. CONCLUSIONS

To maintain operation stability of the intricate modern power grids under various unanticipated events and practical constraints, an SGA-RL method is proposed to perform time-series active power corrective control. Relevant methods are developed to enhance the robustness and efficacy of our control agents, including GATs for better generalization on topologies, MCTS for huge action space reduction, reward-guided differential exploration strategy for deep action evaluation, and long-short term action deployment strategy for comprehensive control effects. Case studies illustrate the capability and great potential of the proposed method for controlling the power system active power considering various contingencies and strict constraints.

Future research will focus on the strengthening of the RL agents' performance based on machine learning and domain knowledge. In addition, as the size of the input graph is associated with power equipment, the training and inference of GATs could be highly influenced in large grids. Therefore, the scalability of our method for bulk systems needs further evaluation as well. Graph formulation and graph pooling will be addressed in future papers as a modification to our method.

REFERENCES

- [1] A. Y. M. Abbas, S. E. G. M. Hassan, and Y. H. Abdelrahman, "Transmission lines overload alleviation by generation rescheduling and load shedding," *Journal of Infrastructure Systems*, vol. 22, no. 4, pp. A4016001, May 2016.
- [2] M. Sahraei-Ardakani and K. W. Hedman, "Day-ahead corrective adjustment of FACTS reactance: a linear programming approach," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2867–2875, Jul. 2016.
- [3] R. Y. Bi, T. Lin, R. S. Chen, J. Ye, X. M. Zhou, and X. L. Xu, "Alleviation of post-contingency overloads by SOCP based corrective control considering TCSC and MTDC," *IET Generation, Transmission & Distribution*, vol. 12, no. 9, pp. 2155–2164, May 2018.
- [4] B. Gou and H. M. Zhang, "Fast real-time corrective control strategy for overload relief in bulk power systems," *IET Generation, Transmission & Distribution*, vol. 7, no. 12, pp. 1508–1515, Dec. 2013.
- [5] E. A. Goldis, P. A. Ruiz, M. C. Caramanis, X. G. Li, C. R. Philbrick, and A. M. Rudkevich, "Shift factor-based SCOPF topology control MIP formulations with substation configurations," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1179–1190, Mar. 2017.
- [6] J. D. Fuller, R. Ramasra, and A. Cha, "Fast heuristics for transmission-line switching," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1377–1386, Aug. 2012.
- [7] T. Ding and C. Y. Zhao, "Robust optimal transmission switching with the consideration of corrective actions for $N - k$ contingencies," *IET Generation, Transmission & Distribution*, vol. 10, no. 13, pp. 3288–3295, Oct. 2016.
- [8] M. Khanabadi, H. Ghasemi, and M. Doostizadeh, "Optimal transmission switching considering voltage security and N-1 contingency analysis," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 542–550, Feb. 2013.
- [9] E. A. Goldis, X. G. Li, M. C. Caramanis, B. Keshavamurthy, M. Patel, A. M. Rudkevich, and P. A. Ruiz, "Applicability of topology control algorithms (TCA) to a real-size power system," in *Proceedings of 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013, pp. 1349–1352.
- [10] M. Abdi-Khorsand, M. Sahraei-Ardakani, and Y. M. Al-Abdullah, "Corrective transmission switching for $N-1-1$ contingency analysis," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1606–1615, Mar. 2017.
- [11] A. G. Bakirtzis and A. P. S. Meliopoulos, "Incorporation of switching operations in power system corrective control computations," *IEEE Transactions on Power Systems*, vol. 2, no. 3, pp. 669–675, Aug. 1987.
- [12] A. A. Mazi, B. F. Wollenberg, and M. H. Hesse, "Corrective control of power system flows by line and bus-bar switching," *IEEE Transactions on Power Systems*, vol. 1, no. 3, pp. 258–264, Aug. 1986.
- [13] W. Shao and V. Vittal, "Corrective switching algorithm for relieving overloads and voltage violations," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1877–1885, Nov. 2005.
- [14] B. Li and G. Sansavini, "Effective multi-objective selection of inter-subnetwork power shifts to mitigate cascading failures," *Electric Power Systems Research*, vol. 134, pp. 114–125, May 2016.
- [15] L. D. Arya and A. Koshti, "Anticipatory load shedding for line overload alleviation using Teaching learning based optimization (TLBO)," *International Journal of Electrical Power & Energy Systems*, vol. 63, pp. 862–877, Dec. 2014.
- [16] M. Y. Li, P. B. Luh, L. D. Michel, Q. C. Zhao, and X. C. Luo, "Corrective line switching with security constraints for the base and contingency cases," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 125–133, Feb. 2012.
- [17] A. S. Korad and K. W. Hedman, "Robust corrective topology control for system reliability," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4042–4051, Nov. 2013.
- [18] A. A. Abou EL Ela, A. Z. El-Din, and S. R. Spea, "Optimal corrective actions for power systems using multiobjective genetic algorithms," in *Proceedings of 42nd International Universities Power Engineering Conference*, Brighton, 2007, pp. 365–376.
- [19] X. P. Li, P. Balasubramanian, M. Sahraei-Ardakani, M. Abdi-Khorsand, K. W. Hedman, and R. Podmore, "Real-time contingency analysis with corrective transmission switching," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2604–2617, Jul. 2017.
- [20] D. Ram, L. Srivastava, M. Pandit, and J. Sharma, "Corrective action planning using RBF neural network," *Applied Soft Computing*, vol. 7, no. 3, pp. 1055–1063, Jun. 2007.
- [21] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: past considerations and perspectives," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6918–6927, Jul. 2017.
- [22] Z. Zhang, D. Zhang and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, Mar. 2020, doi: 10.17775/CSEEJPES.2019.00920.
- [23] F. Ruelens, B. J. Claessens, P. Vrancx, F. Spiessens and G. Deconinck, "Direct load control of thermostatically controlled loads based on sparse observations using deep reinforcement learning," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 4, pp. 423–432, Dec. 2019, doi: 10.17775/CSEEJPES.2019.00590.
- [24] J. J. Duan, D. Shi, R. S. Diao, H. F. Li, Z. W. Wang, B. Zhang, D. S. Bian, and Z. H. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 814–817, Jan. 2020.
- [25] T. Lan, J. J. Duan, B. Zhang, D. Shi, Z. W. Wang, R. S. Diao, and X. H. Zhang, "AI-based autonomous line flow control via topology adjustment for maximizing time-series ATCs," in *Proceedings of the IEEE Power & Energy Society General Meeting (PESGM)*, 2020, pp. 1–5.
- [26] P. D. Xu, Y. Z. Pei, X. H. Zheng, and J. Zhang, "A simulation-constraint graph reinforcement learning method for line flow control," in *Proceedings of the 4th IEEE Conference on Energy Internet & Energy System Integration*, Wuhan, China, 2020, pp. 319–324.
- [27] Universidad Nacional de Colombia. (2020, Oct. 20). L2RPN-NEURIPS-2020 [Online]. Available: <https://github.com/unaioperator/l2rpn-neurips-2020>.
- [28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," arXiv e-prints arXiv: 1710.10903, Oct. 2017. [Online]. Available: <https://arxiv.org/pdf/1710.10903.pdf>.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, New York, USA, 2017, pp. 6000–6010.
- [30] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis,

“Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

- [31] RTE-France. (2021, Aug. 22). Grid2Op [Online]. Available: <https://github.com/rte-france/Grid2Op>.
- [32] BDonnot. (2021, Oct. 8). L2RPN NEURIPS 2020 - robustness track [Online]. Available: <https://competitions.codalab.org/competitions/25426>.



Peidong Xu (S'19) received the M.S. degree in School of Electrical Engineering and Automation from Wuhan University, China, in 2018. He is currently pursuing a Ph.D. degree in School of Electrical Engineering and Automation from Wuhan University. His research interests include power system operation and control, artificial intelligence, and electric vehicles.



Jiajun Duan (S'13–M'18) received the B.S. degree in Power System and its Automation from Sichuan University, Chengdu, China, and M.S. degree in Electrical Engineering at Lehigh University, Bethlehem, PA in 2013 and 2015, respectively, and the Ph.D. degree in Electrical Engineering from Lehigh University in 2018. Currently, he is a Research Scientist in GEIRI North America, San Jose, CA, USA. His research interests include artificial intelligence, power system, power electronics, control systems, and machine learning.



Jun Zhang (M'09–SM'14) received the B.E. and M.E. degrees in Electrical Engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2005, respectively, and the Ph.D. degree in Electrical Engineering from Arizona State University, Phoenix, AZ, USA, in 2008. He is currently a Professor in the School of Electrical Engineering and Automation, Wuhan University. He has authored/coauthored more than 80 peer-reviewed publications. His research expertise is in the areas of complex systems, artificial intelligence, knowledge

automation, and their applications in smart grid.



Yangzhou Pei received the B.S. degree in College of Electrical and Power Engineering, Taiyuan University of Technology in 2019. She is currently pursuing a Ph.D. degree in School of Electrical Engineering and Automation from Wuhan University. Her research interests include artificial intelligence, power system stability control.



Di Shi (M'12–SM'17) received the B.S. degree in Electrical Engineering from Xian Jiaotong University, Xian, China, in 2007, and M.S. and Ph.D. degrees in Electrical Engineering from Arizona State University, Tempe, AZ, USA, in 2009 and 2012, respectively. He currently leads the AI & System Analytics Group at GEIRI North America, San Jose, CA, USA. His research interests include WAMS, Energy storage systems, and renewable integration. He is an Editor of IEEE Transactions on Smart Grid.



Zhiwei Wang (M'16–SM'18) received the B.S. and M.S. degrees in Electrical Engineering from Southeast University, Nanjing, China, in 1988 and 1991, respectively. He is President of GEIRI North America, San Jose, CA, USA. Prior to this assignment, he served as President of State Grid US Representative Office, New York City, from 2013 to 2015, and President of State Grid Wuxi Electric Power Supply Company from 2012–2013. His research interests include power system operation and control, relay protection, power system planning, and WAMS.



Xuzhu Dong (M'02–SM'10) received the Ph.D. degree in High Voltage Engineering from Tsinghua University, Beijing, China, in 1998, and the Ph.D. degree in Electrical Engineering from Virginia Tech University, USA, in 2002. He is currently a Professor in the School of Electrical Engineering and Automation, Wuhan University. His research interests include distribution automation, energy storage, renewable energy and micro-grid.



Yuanzhang Sun (M'99–SM'01) received the Ph.D. degree in Electrical Engineering from Tsinghua University, Beijing, China, in 1988. He is currently a Professor in the School of Electrical Engineering and Automation, Wuhan University, and a Chair Professor in the Department of Electrical Engineering and Vice Director of the State Key Laboratory of Power System Control and Simulation, Tsinghua University. His main research interests are power system dynamics and control, wind power, voltage stability and control, and reliability.