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*, Yuanzhang Sun, *Senior Member*, *IEEE*

***Abstract*—This paper addresses the active power corrective control of modern power systems by adopting deep reinforcement learning. The strategy aims to minimize the joint effect of operation cost and blackout penalty, while robustness and adaptability of the control agent are studied. In Part I of this paper, we consider the robustness case, where the agent is developed to deal with unexpected incidents and guide the stable operation of power grids. A simulation-driven graph attention reinforcement learning (SGA-RL) method is proposed to perform robust active power corrective control. The graph attention networks are introduced to learn the representation of power system states considering topological features. Monte Carlo tree search is utilized to select eligible actions from large action space, including generator redispatch and topology control actions. Finally, driven by simulation, a guided training mechanism and a long-short term action deployment strategy are designed to help the agent better evaluate the action set while training and operate more stably while deploying. The effectiveness of the proposed method is demonstrated in "2020 Learning to Run a Power Network - Neurips Track 1" global competition and relevant cases. In Part II of this paper, we address the adaptability case, where the agent is established to adapt to the grid with an increasing share of renewable energies over years.**

***Index Terms*—Active power corrective control; Deep reinforcement learning; Simulation-driven; Graph attention networks.**

NOMENCLATURE

$\mathbf{B}$	Connected buses of power equipment in substations.
c	Constant to control the searching depth.
$C(\cdot)$	Function associated with the adjustment cost.
$C_{net}(t)$	Network loss cost.
$E_{loss}(t)$	Energy loss at time t when a blackout occurs.
$\boldsymbol{h_i}$	Transformed features of the ith node.
K	Number of independent attention mechanisms.
$n(s_t)$	Simulation number of node s_t.
N_C	Number of allowed topological actions.
N_L	Number of lines.
N_S	Total simulation number.
N_i	Neighborhood node set of the ith node.
$p(t)$	Marginal price.
$\Delta\mathbf{P_G}$, $\Delta\mathbf{P_L}$	Amount of generator redispatch and load shedding in the power system, respectively.
$\mathbf{P_{G\max}}$, $\mathbf{P_{G\min}}$	Upper and lower bounds of generator outputs.
$\mathbf{P_{redisp}}$	Redispatch amount of each generator.
$\mathbf{P}$, $\mathbf{Q}$	Power status of power equipment.
$\mathbf{R_{up}}$, $\mathbf{R_{down}}$	Upper and lower limits of generator ramping capabilities.
T	Length of the control period
$T_{s,end}$	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.
$\mathbf{W}^k$	kth weight matrix.
X_{line}, X_{bus}	Number of line switching actions, bus-bar switching actions.
α_{ij}^k	Normalized attention coefficients computed by the kth attention mechanism.
α, β	Penalty factors of overload and heavy load, respectively.
π	Shared attentional mechanism.
$\boldsymbol{\rho}$	Load ratio of each powerline.
σ	Non-linear activation function.
Ω	Predefined threshold of load ratio.

This work was supported by the National Key R&D Program of China under Grant 2018AAA0101504. (Corresponding author: Jun Zhang.)

P. Xu, J. Zhang, Y. Pei, X. Dong and Y. Sun are with the School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China (e-mail: xupd@whu.edu.cn; jun.zhang.ee@whu.edu.cn; 2019102070001@whu.edu.cn; dongxz@whu.edu.cn; yzsun@mail.tsinghua.edu.cn).

J. Duan, D. Shi and Z. Wang are with GEIRI North America, San Jose, CA 95134, USA. (e-mail: jiajun.duan@geirina.net; di.shi@geirina.net; zhiwei.wang@geirina.net).

DOI: 10.17775/CSEEJPES.2020.07090

I. INTRODUCTION

WITH wide adoption of ultra-high voltage transmission technology and renewable energy integration, the power grids often operate under large scale and long-distance power

exchange conditions, which makes the task of maintaining the stable operation of power systems even more challenging. When power line overloads occur, actions must be taken to eliminate or alleviate the overloads, otherwise succeeding cascading failure may be triggered. Therefore, it is of great importance to study fast and efficient active power corrective control techniques for the safe and stable operation of the power network.

Conventional corrective control actions mainly focus on the generator redispatch or load shedding, where the node injections are optimized to adjust line power flow [1]. With the development of modern power systems, more control measures can be introduced to deal with line overloads. In [2], flexible AC transmission system (FACTS) devices are adopted to perform day-ahead corrective control in security-constraint unit commitment (SCUC) problems. Multi-terminal direct current (MTDC) power and series capacitors (SC) switching are also utilized to relieve line overloads [3-4]. With the promising flexibility, topology control, such as transmission switching (TS) and substation configurations, is considered an effective power flow control technology and has achieved good cost reduction in economic dispatch and transmission congestion management [5-9]. Besides, by adopting corrective TS, the ramping capability of generators can be preserved so the system can be more robust [10]. In [11-13], topology actions are proved to be reliable in corrective overload control. Thus, line and bus-bar switching can be utilized as feasible corrective control actions in addition to conventional measures.

At present, common corrective control methods can be categorized into sensitivity-based methods and optimization-based methods. The first category of methods utilizes sensitivity factors to quantify the effects of certain measures on the overloaded lines and selects high-sensitivity measures to handle the overloads, thus is computation-efficient and easy to implement, but may suffer from low accuracy due to the adoption of direct current (DC) power flow method [14]. Thus, sensitivity-based methods are usually combined with other methods. In [15], sensitivity factors are introduced to select the most effective load shed buses for further optimization. In the optimization-based methods, corrective control actions are selected by solving optimization models. By formulating the optimal corrective line switching problem as a constraint satisfaction problem, constraint programming and tree search are adopted to provide a solution [16]. In [17], the robust corrective switching problem is formulated as a mixed integer programming (MIP) problem and robust optimization is utilized to deal with uncertainties. Heuristic methods are also introduced to obtain optimal solutions [18-19]. The optimization-based methods can adapt to the complex constraints of power systems but can be computationally expensive. In [20], an approach based on radial basis function neural network (RBFN) is proposed to seek an efficient approach to line overload alleviation.

Recently, with the rapid development and wide application of reinforcement learning (RL), RL is considered as a viable solution to many decision and control problems across different time scales and electrical system states [21]. Thus, RL can be considered and adopted to perform time-series corrective control for the stable operation of the complex power system. In our previous work, the deep reinforcement learning (DRL) method is utilized to perform autonomous voltage control and line flow control [22-24].

Therefore, based on our previous work, a simulation-driven reinforcement learning method combined with graph neural networks is presented in this paper for active power corrective control. Topology actions and generator redispatch are combined to provide comprehensive corrective measures for the DRL agent. Graph attention networks (GATs) and Monte Carlo tree search (MCTS) are adopted to tackle the two challenges of complex observation space and action space in our problem, respectively. Sub-agents are established to perform parallel training and collaborative deployment based on different action sub-space. A reward-guided deep exploration mechanism is developed to assist sub-agents to learn effective correction actions. 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 proposed method is eventually proved effective in a global power system AI competition and relevant cases.

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

II. Problem Formulation

A. Objective and Constraints

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

$$\min C(|\Delta \mathbf{P_G}|, \Delta \mathbf{P_L}) \tag{1}$$

where $C(\cdot)$ denotes the function associated with the adjustment cost, and $\Delta \mathbf{P_G}$, $\Delta \mathbf{P_L}$ are the amount of generator redispatch and load shedding in the power system, respectively. Note that here we assume topological changes are cost-free.

In addition to conventional power system constraints, the number of topological control actions and redispatch amounts of generators should also be restricted for the minimum disturbance to the power system.

$$X_{line} + X_{bus} \le N_C \tag{2}$$

$$\begin{cases} |\Delta \mathbf{P_G}| \le \min(\mathbf{P_{G\max}} - \mathbf{P_G}, \mathbf{R_{up}}) \\ |\Delta \mathbf{P_G}| \le \min(\mathbf{P_G} - \mathbf{P_{G\min}}, \mathbf{R_{down}}) \end{cases} \tag{3}$$

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

As the active power correction aims to alleviate or eliminate line overloads, to prevent cascading failures and maintain stable operation, in this paper, the problem is extended to a

continuous power system real-time control problem. Based on the well-planned dispatch schedule, corrective actions are taken when normal overloads or post-contingency overloads occur, "do-nothing" action is carried out when no line violates the constraints. The objective is then transformed to maintain the system operation while minimizing the overall cost during the entire period as

$$\min \sum_{t=0}^{T}\left[C\left(\left|\Delta \mathbf{P}_{\mathbf{G}}\left(t\right)\right|, \Delta \mathbf{P}_{\mathbf{L}}\left(t\right), t\right)+C_{net}\left(t\right)+E_{loss}\left(t\right) p\left(t\right)\right] \quad (4)$$

where T denotes the length of the control period, $C_{net}\left(t\right)$ is the network loss cost, which can reflect the economic influence of corrective actions. $E_{loss}\left(t\right)$ is the energy loss at time t when a blackout occurs, and $p(t)$ represents the marginal price. By combining operation cost and blackout penalty, the optimal control strategy can maintain the power grid stable by less costly corrective actions.

Finally, except the above constraints, reaction time, i.e., time to react to an overloaded line before it is disconnected by relay protectors, and recovery time, i.e., the required time to reactivate the power equipment after its activation, are introduced to enhance the practicability 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 $M=\left\{S, A, P, R, \gamma\right\}$. Among these elements, γ denotes a discounting factor, and P is the transition probability matrix. Other elements in our question are elaborated as follows.

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

$$s_t=\left(\mathbf{P}, \mathbf{Q}, \mathbf{B}, \mathbf{P}_{\mathbf{redisp}}, \boldsymbol{\rho}\right) \quad (5)$$

where $\mathbf{P}$ and $\mathbf{Q}$ denote the power status of power equipment, including power outputs of generators, consumption value of loads, and power flow at both ends of lines. $\mathbf{B}$ represents the connected buses of power equipment in substations. $\mathbf{P}_{\mathbf{redisp}}$ is the redispatch amount of each generator. $\boldsymbol{\rho}$ denotes the load ratio of each powerline, i.e., the current flow divided by the thermal limit of each powerline.

Action space A : The action space consists of generator redispatch, line switching, bus-bar switching, and do-nothing. The line switching changes the state of power grids by reconnecting/disconnecting a powerline. The bus-bar switching involves the bus selection of elements connected to the bus-bar, each element can be switched onto either of the buses of the bus-bar to control the power system.

Reward function R : As our primary mission is to maintain stable operation of the power system under load and renewable energy generation fluctuation, maintenance, and contingencies, so the residual capacity of the grid should be optimized to improve the grid flexibility in the presence of unexpected events. Besides, since loaded lines and overloaded lines may become bottlenecks to power transmission, their appearance should be considered as a penalty on the flexibility of the system. Thus, performance at time t can be formulated as [25]:

$$\begin{aligned} o_t=\sum_{i=1}^{N_L}\Big[& \max\left(0,\left(1-\rho_i^2\right)\right)-\alpha\cdot\max\left(0,\rho_i-1\right) \\ & -\beta\cdot\max\left(0,\rho_i-0.9\right)\Big] \end{aligned} \quad (6)$$

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

In this problem, the system is controlled in a time-series manner, and we expect the power system can survive longer. Therefore, if the DRL agent is rewarded by both historical and current performance, the agent will be motivated to select more efficient actions to avoid damaging system stability maintained from the past to the present time step. Hence, 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

Recently, reinforcement learning has demonstrated its great potential in many areas for complex decision-making problems. When performing corrective control in large power systems with unexpected incidents by generator redispatch and topological adjustments, there are several important issues to be addressed: 1) Topological changes due to random events; 2) high dimensions of topological action space; 3) high stability requirement.

To solve these issues, in this paper, a simulation-driven graph attention reinforcement learning (SGA-RL) method is developed, the architecture is shown in Fig.1. In this architecture, graph attention networks are introduced to learn the power grid state representation and enhance the agent's generalization ability to unpredictable topological changes. Then Monte Carlo tree search is used to limit the action space. Finally, a reward-guided training mechanism and a long-short term action deployment strategy are designed to achieve effective corrective control strategy learning and stable application, respectively. To prevent the power system from violating the constraints, an early warning mechanism is introduced [23], i.e., when the maximum load ratio in the simulation exceeds the predefined threshold Ω, e.g., 0.98, the agent is also activated to provide control decisions, to protect the system in a preventive and corrective way. Double dueling deep Q-network is also utilized as the basic DRL framework.

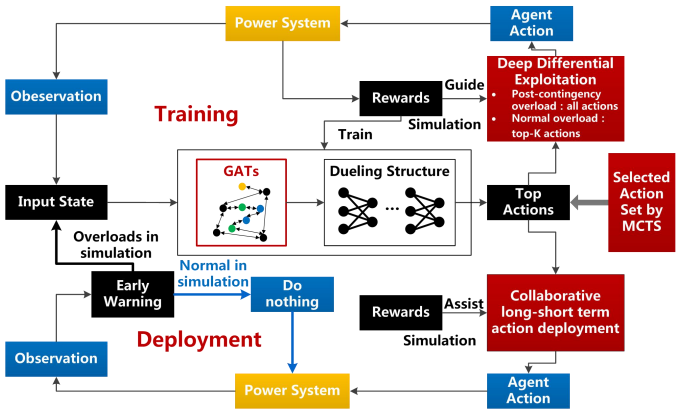


Fig. 1. The architecture of the proposed method.

A. Graph Attention Networks

Graph neural networks are utilized to extract features from power grid states. The state is transformed into graph structure data as in [24]. The adjacency matrix $\mathbf{A}$ can be established based on the connections between power equipment, i.e., ends of lines, generators, and loads. The feature matrix $\mathbf{X}$ is established by combining the common features and unique features of different equipment. An example is shown in Fig. 2 and (8).

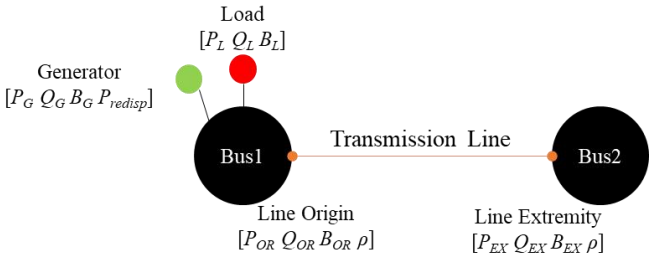


Fig. 2. Example for the establishment of adjacency matrix and feature matrix.

$$\begin{cases} \mathbf{A} = \begin{matrix} \\ \text{Origin} \\ \text{Extremity} \\ \text{Load} \\ \text{Generator} \end{matrix} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \\ \mathbf{X} = \begin{matrix} \\ \text{Origin} \\ \text{Extremity} \\ \text{Load} \\ \text{Generator} \end{matrix} \begin{bmatrix} P_{OR} & Q_{OR} & B_{OR} & 0 & \rho \\ P_{EX} & Q_{EX} & B_{EX} & 0 & \rho \\ P_L & Q_L & B_L & 0 & 0 \\ P_G & Q_G & B_G & P_{resisp} & 0 \end{bmatrix} \end{cases} \tag{8}$$

where P_{OR} and Q_{OR} denote the power flow at the origin of the transmission line, P_{EX} and Q_{EX} represent the power flow at the extremity of the transmission line, P_L and Q_L represent the consumption value of the load, P_G and Q_G denote the power output of the generator. B_{OR}, B_{EX}, B_L, B_G represent the connected buses of above power equipment in the substation. P_{resisp} is the redispatch amount of generator, and ρ is the load ratio of the transmission line.

In our problem, when the grid undergoes maintenance or "N-1" events, the topologies can be quite different from normal ones. Therefore, the DRL agent should possess sufficient generalization ability to decide under unseen topologies caused by random events and to have better performance on various operation scenarios.

Thus, graph attention networks are utilized to perform representation learning on 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 aggregated solely based on the information from its neighboring nodes. Hence, models utilizing GATs can be generalized to analyze different structures. The graph attention layer with multi-head attention for transformed features of the *i*th node $\boldsymbol{h}_i$ is defined in (9) [26]:

$$\boldsymbol{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k \mathbf{W}^k \boldsymbol{h}_j\right) \tag{9}$$

where σ is the non-linear activation function, N_i is the neighborhood node set of the *i*th node, K denotes the number of independent attention mechanisms, α_{ij}^k represents the normalized attention coefficients computed by the *k*th attention mechanism [27], 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 perform self-attention. The importance of neighboring *j*th node's features to *i*th node can be calculated based on the shared attentional mechanism: $\pi\left(\mathbf{W}^k\boldsymbol{h}_i, \mathbf{W}^k\boldsymbol{h}_j\right)$. By adopting the softmax function α_{ij}^k can be obtained. As we combine the normalized attention coefficients of K attention mechanisms in from of (9), the multi-head attention is realized and the aggregated features of *i*th node is produced.

B. Monte Carlo Tree Search

In large scale power systems, there will be a large number of available corrective actions due to the discreteness and combination of 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 difficult or even impossible to train.

To reduce the dimension of the action space, Monte Carlo tree search is adopted to select the effective actions by sampling and evaluating from the entire action space. Typical scenarios are chosen to form the searching set, in each scenario, "do-nothing" or "auto-reclosing" is performed unless overload occurs. The state with line overload is considered as the root node, and the selection-expansion-simulation-backpropagation process can be performed to update the tree [28].

To serve the purpose of finding effective corrective actions, the leaf nodes of the tree denote the next overload states after applying feasible actions in the previous overload states, the edges represent the corrective actions a_t. Each node s_t of the tree stores its state evaluation value $W(s_t)$ and simulation number $n(s_t)$. $W(s_t)$ 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):

$$\begin{cases} u\left(n\left(s_t\right)\right)=\dfrac{T_{s,end}-t}{T-t} \\ W\left(s_t\right)_{n(s_t)}=W\left(s_t\right)_{n(s_t)-1}+u\left(n\left(s_t\right)\right) \end{cases} \tag{10}$$

where $T_{s,end}$ 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 are performed for the current overload state to form the effective action set. After the enumeration, at the subsequent MCTS process, action selection will follow 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.

$$score=\frac{W\left(s_t\right)}{n\left(s_t\right)}+c\sqrt{\frac{\ln N_S}{n\left(s_t\right)}} \tag{11}$$

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

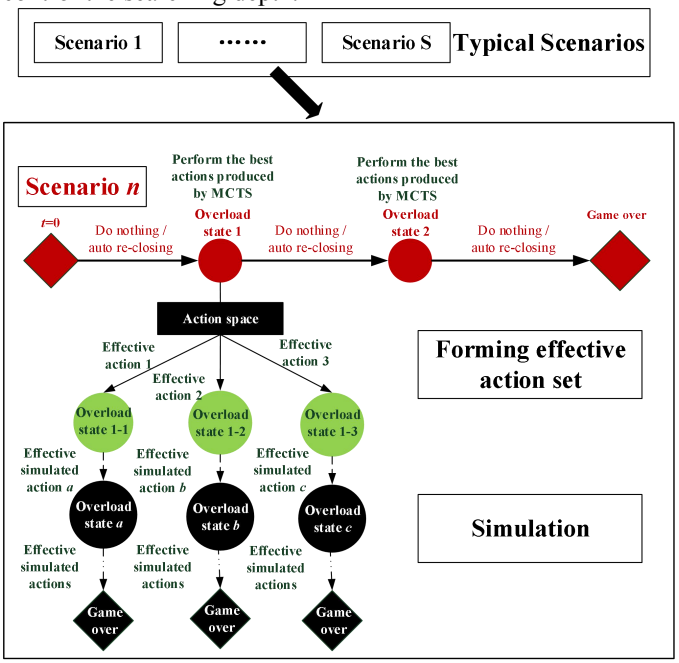


Fig. 3. Monte Carlo tree search in our problem.

C. Reward-guided Training Mechanism

After feasible actions are identified, the action space will still have high dimensions for bulk power systems or complex substation configurations. To further improve 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. Besides, at each time step, the prior auto-reclosing check is carried out before 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 unexpected events and relieve the control difficulty for the sub-agent.

As line overloads can lead to cascading failures, so the effectiveness of corrective actions is strongly required. Thus, in overload states, δ actions with the highest Q-values are selected and checked by the simulation system in descending order. The feasible action with highest simulated reward is chosen as the final decision. When the grid is in a post-contingency state, the loaded lines or overloaded lines can be great risks to the grid stability, so here δ equals to the number of entire sub-agent action space, i.e., enumerating all actions to find an action with the best reward. The reward-guided differential exploration strategy will help the sub-agent evaluate the action set more precisely, especially in a post-contingency state, to perform valid actions when deploying.

In most operational states, "do-nothing" is carried out by the DRL model owing to the pre-dispatching schedule. When adopting experience replay, many "do-nothing" transitions will be stored in replay buffer, leading to the low sample frequency of other useful transitions to perform experience replay, which can obstruct the proper evaluation of corrective control actions. In this paper, a balanced replay buffer is designed: The "do-nothing" transitions are stored in the buffer with a small 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 I.

Algorithm I Simulation-driven Graph DQN Training Method

Initialize weights of Graph DQN θ
Initialize Memory Buffer Δ
for episode = 1 to I **do**
  Reset the environment
  for t = 1 to T **do**
    Collect and transform the observation s_t
    Perform do-nothing action a^0 in the simulation system, check the danger flag (True for overload or system failure, False for normal state)
    if the danger flag is True **then**
      Select δ top actions according to $Q\left(\cdot|s_t\right)$ of the agent:

$$\delta=\begin{cases}|A| & \text{if post-contingency state} \\ V & \text{otherwise}\end{cases}$$

      Validate the actions in the simulation system and choose the feasible action with the best reward as the output action a_t
    else then
      select do-nothing action a^0 as the output action a_t
    end if
    Execute action a_t in the environment and obtain next state s_{t+1}, reward r_t, and d_t
    Store the transition (s_t, a_t, r_t, s_{t+1}, d_t) in Δ, if $a_t = a^0$, save with a small probability p_s, if d_t is True, save multiple times
    Sample a batch of transitions from Δ using importance sampling
    Calculating the target Q-values:

$$y_i=\begin{cases}r_i & \text{if } d_t \text{ is True} \\ r_i+\gamma\max_{a'}Q\left(s_{t+1},a';\theta^-\right) & \text{otherwise}\end{cases}$$

    Update Q-network by losses:

$$L_i\left(\theta\right)=\left(y_i-Q\left(s_t,a_t;\theta\right)\right)^2$$

    Hard copy main network weights θ to the target network regularly
    Update state $s_t = s_{t+1}$
  end for
end for

D. Long-Short Term Action Deployment Strategy

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

application should be both considered.

Based on the collaboration of all trained sub-agents, a long-short term action deployment strategy is proposed. 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. Note that k varies in different types of overloads.

In 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):

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

The base action is then updated by other sub-agents according to simulated reward, to promote the better short-term control effect of final action. In one sub-agent, its verified top actions are compared with base action by simulated rewards. When a top action prevails the base action by 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 way, the output control action will have a good immediate control effect while maintaining long-term benefit.

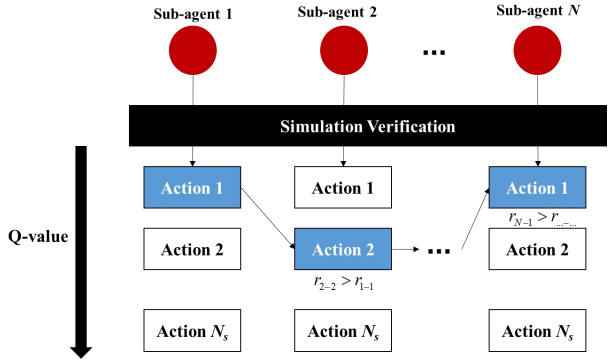


Fig. 4. Long-short term action deployment strategy.

IV. Case Studies

A. Experimental Setup

The open-source platform Grid2Op [29] is utilized 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 covering 12 months of 48 years. Each scenario contains data for 28 continuous days in 5-minute resolution. Attacks on random transmission lines will happen every day at different times, i.e., "N-1" event occurs such as that the corresponding lines will suffer outage for an uncertain duration. Meanwhile, the maintenance of lines also exists. In the platform, 24 one-week validation scenarios that have not been seen before are provided to evaluate the trained agents. These scenarios are selected from each month of the year with special considerations [30].

The number of topological control actions in our research is restricted to 1, i.e., $X_{line} + X_{bus} \le 1$, to avoid creating 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 timesteps, i.e., 15 minutes.

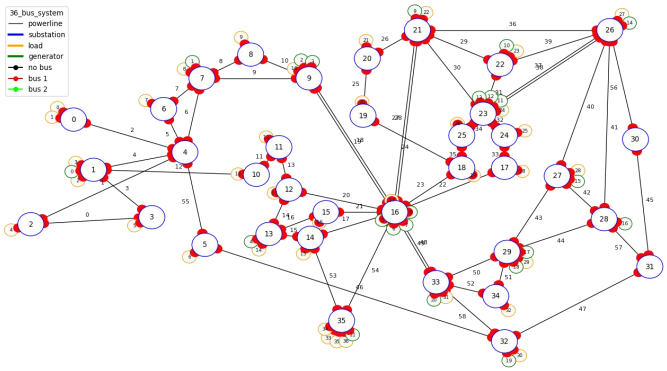


Fig. 5. The target 36-bus system.

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

B. Performance of the Proposed Method

The proposed simulation-driven graph reinforcement learning method is adopted to train the DRL agents for corrective control. After applying Monte Carlo tree search, around 1500 eligible actions are selected. The entire action space is divided into 4 parts with each learned 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 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 verified. The "do-nothing" transitions are stored with a probability equal to 0.05.

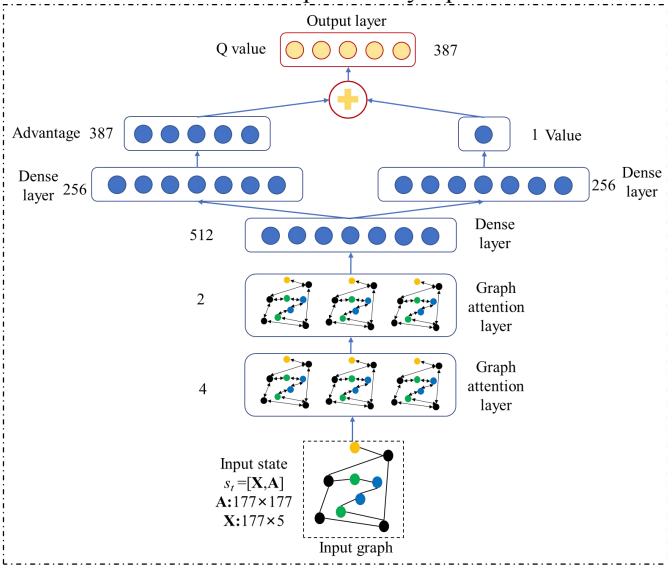


Fig. 6. The architecture of graph DQN.

During the training, the system operating duration of each episode, i.e., the number of timesteps the power system maintains stable before the blackout, is also recorded. The cumulative rewards curve and system operating duration curve

of a sub-agent during training are shown in Fig. 7. The cumulative rewards and system operating duration are averaged and smoothed.

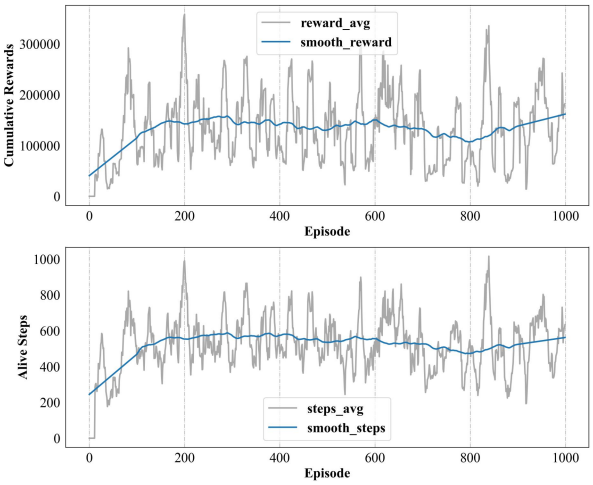
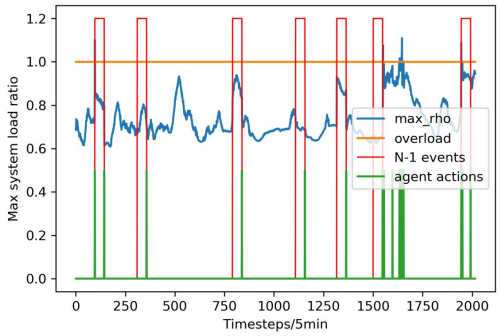


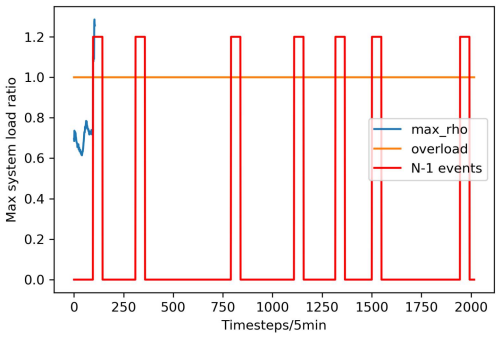
Fig. 7. The training process of the graph DQN agent.

Considering the diversity of training scenarios, e.g., different generation and load status, stochastic "N-1" events, and reward-guided exploration process of sub-agent, fluctuations are caused in the cumulative rewards and system operating duration. However, according to the smoothed curves in Fig. 7, the sub-agent tends to improve its performance during the first 200 episodes and maintains the performance for the rest of the training period.

After 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 Ω=0.98 . A local validation scenario is selected to evaluate the graph DQN agents, while a "do-nothing" agent is also evaluated to serve as a baseline. The overall system operation status and agents' decisions are shown in Fig. 8.



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



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

Fig. 8. System operation status comparison.

During the one-week scenario of 2016 time steps, 7 "N-1" events occur, the starting timesteps are 96, 310, 791, 1109, 1317, 1501, 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., fails during the first "N-1" event. Specifically, under the control of graph DQN model, 5 overloads occur during the operation, 2 of them are triggered by the first and last "N-1" events, respectively, while the rests 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 show 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, the overload occurs. Our model performs a bus-bar switching action on bus 28: Generator 16, the extremity of line 42 and line 57 are connected on bus 28-1, and the extremity of line 41 and origin of line 44 are assigned on 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 find that the overload does not disappear and can lead to system failure if no control measure is taken. Thus, the corrective action made by our model is necessary and effective.

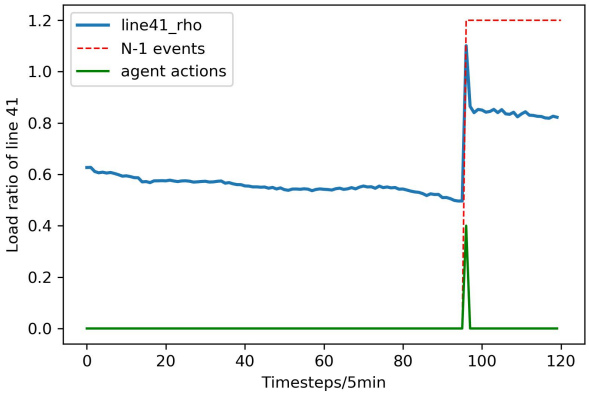


Fig. 9. Status of line 41 during the first "N-1" event.

To further evaluate 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. For more readable, 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 scenarios successfully with topological actions and minimal network loss. The trained DRL model is submitted to the Robustness Track of L2RPN competition, another 24 unseen scenarios are provided by the competition host to test our model, the results are shown in Table I.

TABLE I
PERFORMANCE OF PROPOSED MODEL ON VALIDATION/TEST SCENARIOS

	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 4-day long stable operation in average under unexpected incidents. Besides, the performance on validation scenarios and test scenarios have barely any difference, the latter even slightly outperforms the former, indicating the generalization ability of our model. The average decision time for each time step of our model is around 35 ms. The results show the potential of the proposed method in alleviating overloads in complex power systems.

C. *Effectiveness of Graph Attention Networks*

To check the feasibility of adopting GATs, a GCN-based model is utilized for comparation. The two models differ from each other in the use of graph neural networks. Two graph convolution layers are utilized 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.

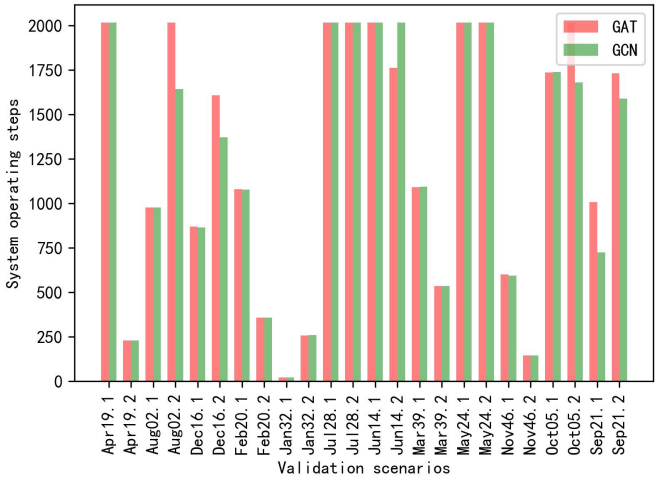


Fig. 10. System operating duration under different models.

TABLE II
PERFORMANCE COMPARISON BETWEEN DIFFERENT MODELS

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

From Fig. 10 and Table II, we can find that both models perform well on the validation set, and the GAT model has a slight advantage in supporting the power system's continuous operation, achieving better benefits in stability and economy. Thus, it is recommended to introduce GATs in the reinforcement learning model if the power grid may undergo large topological changes.

D. *Effectiveness of Long-Short Term Action Deployment Strategy*

Based on the trained sub-agents, a fully reward-guided action deployment strategy and an enumeration strategy are adopted to compare the performance of our long-short term action strategy. In the fully reward-guided strategy, 100 and all top actions are simulated in each sub-agent to find the action with the best reward, to eliminate normal overload and post-contingency overload, respectively. In the enumeration strategy, all the sub-agents' actions are enumerated to find the best action with the largest immediate reward.

We first apply the two strategies 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.

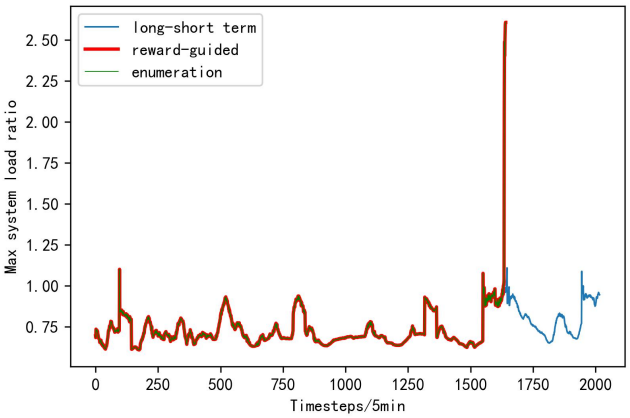


Fig. 11. System operating status under different strategies.

TABLE III
OPERATING STATUS UNDER DIFFERENT STRATEGIES

	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 cannot handle the overload in the scenario properly as our proposed long-short term method does, which leads to severe overload and eventually cause blackout at time step 1640. Besides, fewer control actions are conducted under our strategy during the same period, showing the ability of stable and effective power grid control by considering long-term benefits and short-term rewards.

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

TABLE IV
PERFORMANCE COMPARISON BETWEEN DIFFERENT STRATEGIES

	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 prevail 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 complexity 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 evaluate the relevant performance. The results from validation scenarios are shown in Table V.

TABLE V
PERFORMANCE COMPARISON BETWEEN DIFFERENT THRESHOLDS

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

As shown in Table V, the different thresholds to trigger the agents can have a relatively high influence on the performance. If the control strategy is too preventive, unnecessary actions may disturb the power system before events occur. However, a high threshold may prevent the agents from taking action in time due to the strict constraints. Thus, a proper threshold must be selected according to the system characteristics and constraints.

V. CONCLUSIONS

To maintain stable operation of complex power girds under various unexpected events and practical constraints, a SGA-RL method is proposed to perform time-series active power corrective control. Relevant measures are developed to improve the robustness and efficiency 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 demonstrate the capability and great potential of the proposed method in controlling the power system active power considering contingencies and strict constraints.

Future work will focus on the improvement of RL agents' performance from machine learning and domain knowledge. In addition, as the size of input graph is associated with power equipment, the training and inference of GATs may be impacted in large grids. Hence, the scalability of our method on bulk systems should also be further studied. Graph formulation and graph pooling will be addressed to modify our method.

REFERENCES

[1] A. Y. M. Abbas, S. E. G. M. Hassan, and Y. H. Abdelrahim, "Transmission lines overload alleviation by generation rescheduling and load shedding," *J. Infrastruct. Syst.*, 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 Trans. Power Syst.*, vol. 31, no. 4, pp. 2867-2875, Jul. 2016.

[3] R. Bi, T. Lin, R. Chen, J. Ye, X. Zhou and X. 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. 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. Li, C. R. Philbrick and A. M. Rudkevich, "Shift Factor-Based SCOPF Topology Control MIP Formulations with Substation Configurations," *IEEE Trans. Power Syst.*, 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 Trans. Power Syst.*, vol. 27, no. 3, pp. 1377-1386, Aug. 2012.

[7] T. Ding and C. 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 Trans. Power Syst.*, vol. 28, no. 1, pp. 542-550, Feb. 2013.

[9] E. A. Goldis, X. 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 *Proc. 51st Annu. Allerton Conf.*, pp. 1349-1352, Oct. 2013.

[10] M. Abdi-Khorsand, M. Sahraei-Ardakani and Y. M. Al-Abdullah, "Corrective Transmission Switching for N-1-1 Contingency Analysis,"" *IEEE Trans. Power Syst.*, 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 Trans. Power Syst.*, 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 Trans. Power Syst.*, 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 Trans. Power Syst.*, 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," *Electr. Power Syst. Res.*, 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)," *Int. J. Electr. Power Energy Syst.*, vol. 63, pp. 862-877, dec. 2014.

[16] M. Li, P. B. Luh, L. D. Michel, Q. Zhao and X. Luo, "Corrective Line Switching With Security Constraints for the Base and Contingency Cases," *IEEE Trans. Power Syst.*, 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 Trans. Power Syst.*, 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 *Proc. 42nd Int. UPEC*, Brighton, 2007, pp. 365-376.

[19] X. Li, P. Balasubramanian, M. Sahraei-Ardakani, M. Abdi-Khorsand, K. W. Hedman and R. Podmore, "Real-Time Contingency Analysis With Corrective Transmission Switching," *IEEE Trans. Power Syst.*, 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," *Appl. Soft Comput.*, 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," in *20th IFAC World Congress*, 2017, pp. 6918-6927.

[22] J. Duan. D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian and Z. Yi, "Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814-817, Jan. 2020.

[23] T. Lan, J. Duan, B. Zhang, D. Shi, Z. Wang, R. Diao, and X. Zhang, "AI-based autonomous line flow control via topology adjustment for maximizing time-series atcs," *arXiv e-prints*, Nov. 2019.

[24] P. Xu, Y. Pei, X. Zheng and J. Zhang, "A Simulation-Constraint Graph Reinforcement Learning Method for Line Flow Control," in *the 4th IEEE*

Conf. Energy Internet & Energy System Integration, Wuhan, 2020, to be published.

[25] Universidad Nacional de Colombia, L2RPN-NEURIPS-2020 [Online]. Available: https://github.com/unaioperator/l2rpn-neurips-2020.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, "Graph Attention Networks," *arXiv e-prints*, Oct. 2017.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," in *Proc. the 31st Int. Conf. Neural Information Processing Systems (NIPS'17)*, New York, USA, 2017, pp. 6000–6010.

[28] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. 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, pp. 484–489, Jan. 2016.

[29] RTE-France, Grid2Op [Online]. Available: https://github.com/rte-france/Grid2Op.

[30] RTE-France, L2RPN NEURIPS 2020 - Robustness Track [Online]. Available: https://competitions.codalab.org/competitions/25426.