Graph Attention Network Based Deep Reinforcement Learning for Voltage/var Control of Topologically Variable Power System

Xiaofei Liu, *Student Member, IEEE*, Pei Zhang, *Fellow, IEEE*, Hua Xie, *Member, IEEE*, Xuegang Lu, Xiangyu Wu, *Member, IEEE*, and Zhao Liu, *Member, IEEE*

***Abstract*—The high proportion of renewable energy integration and the dynamic changes in grid topology necessitate the enhancement of voltage/var control (VVC) to manage voltage fluctuations more rapidly. Traditional model-based control algorithms are becoming increasingly incompetent for VVC due to their high model dependence and slow online computation speed. To alleviate these issues, this paper introduces a graph attention network (GAT) based deep reinforcement learning for VVC of topologically variable power system. Firstly, combining the physical information of the actual power grid, a physics-informed GAT is proposed and embedded into the proximal policy optimization (PPO) algorithm. The GAT-PPO algorithm can capture topological and spatial correlations among the node features to tackle topology changes. To address the slow training, the ReliefF-S algorithm identifies critical state variables, significantly reducing the dimensionality of state space. Then, the training samples retained in the experience buffer are designed to mitigate the sparse reward issue. Finally, the validation on the modified IEEE 39-bus system and an actual power grid demonstrates superior performance of the proposed algorithm compared with state-of-the-art algorithms, including PPO algorithm and twin delayed deep deterministic policy gradient (TD3) algorithm. The proposed algorithm exhibits enhanced convergence during training, faster solution speed, and improved VVC performance, even in scenarios involving grid topology changes and increased renewable energy integration. Meanwhile, in the adopted cases, the network loss is reduced by 6.9%, 10.8%, and 7.7%, respectively, demonstrating favorable economic outcomes.**

***Index Terms*—Voltage/var control, grid topology, renewable energy, graph attention network, deep reinforcement learning.**

I. Introduction

With the increasing penetrations of renewable energy resources such as wind and solar, the randomness and fluctuation of these resources bring more uncertainty to power systems, resulting in more frequent and severe voltage fluctuations [1], [2], which poses significant challenges for intraday real-time voltage optimization and control. If the voltage cannot be regulated to the normal range in a timely manner by the reactive power compensation devices, the voltage may violate constraints or collapse, resulting in catastrophic accidents [3], [4]. Therefore, it is necessary to study the voltage/var control (VVC) methods in adapting to the uncertainties of wind power and load demand.

Many studies have been conducted to solve the voltage or reactive power optimization and control problem considering the uncertainties of renewable energy generation and load demand [5] - [7]. The commonly used solution methods are mainly divided into four categories: stochastic programming (SP), robust optimization (RO), interval programming (IP), and deep reinforcement learning (DRL). In SP, uncertain variables are expressed by specific probability density as different scenarios [8]. In [9], a two-stage SP model considering the uncertainties of renewable energy generation and load demand is proposed to achieve voltage optimization and control. However, the SP may result in operation constraint violations and a high computation burden. RO deals with the uncertainty of interval uncertain sets [10]. Reference [11] proposes a distributed adaptive robust VVC method to ensure the operation constraints are satisfied. However, RO is only applicable to convex models. In addition, it may neglect the economic objectives. IP regards the uncertain variables of the reactive power optimization model as an interval value, and then formulates a reactive power optimization model considering interval uncertainty [12]. In [13], the uncertain values are expressed as different intervals, and a reactive power optimization model with interval uncertainties is established. Then, the optimization strategy for satisfying the voltage constraints is obtained. However, the performance of IP is directly affected by the formulated uncertainty model.

Unlike SP, RO, and IP, DRL is an artificial intelligence algorithm that does not rely on an accurate physical model. A

Manuscript received: September 27, 2023; revised: March 19, 2024; accepted: June 26, 2024. Date of CrossCheck: June 26, 2024. Date of online publication: October 4, 2024.

This work was supported by China Southern Power Grid Co., Ltd. Yunnan Electric Power Dispatching Control Center (Cyber-physical-based “Source-network-load-storage” Coordinated Dispatch and Control Technologies and Application System Development, sub-project YNKJXM20222463). Many thanks to revision suggestions provided by Zhenhuan Ding from Anhui University, Hefei, China.

X. Liu, P. Zhang (corresponding author), H. Xie, X. Wu, and Z. Liu are with the School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China, and P. Zhang is also with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: 19117015@bjtu.edu.cn; peizhang166@qq.com; hxie@bjtu.edu.cn; wuxiangyu@bjtu.edu.cn; liuzhaol@bjtu.edu.cn).

X. Lu is with Yunnan Electric Power Dispatching and Control Center, Kunming 650011, China (e-mail: 545315893@qq.com).

DOI: 10.35833/MPCE.2023.000712

VVC method based on the deep deterministic policy gradient (DDPG) algorithm is proposed in [14], which verifies the applicability of DDPG under the uncertainty of renewable energy output. However, the stability and convergence speed of DDPG are worse than the proximal policy optimization. An off-policy DRL algorithm based on soft actor-critic (SAC) is proposed in [15], which can achieve VVC with renewable energy resources. However, all samples in the experience buffer are randomly sampled with equal probability, which may result in unstable and low efficient training process. Reference [16] proposes a voltage control method based on deep meta-reinforcement learning to improve the adaptability of DRL algorithm to new grid operation conditions and system parameters. However, the proposed method suffers the sparse reward problem, which can lead to slow training. To cope with the impact of grid topology changes on VVC, [17] proposes a simplified DRL algorithm based on the side-tuning transfer learning algorithm. The proposed algorithm improves the adaptability to different grid topologies while achieving voltage control. However, the proposed algorithm can only be applied to topologies similar to the training topology. A graph convolutional network (GCN)-based DRL algorithm is developed in [18], which is used to deal with the voltage control issue when the topology of the power system changes. However, the importance difference in each node in the power system is not considered. It is well known that wind turbine nodes or load nodes with heavy loads are more likely to cause voltage problems. Therefore, these nodes should be given a higher priority for voltage regulation. In summary, existing DRL-based solutions show two significant problems as follows. Firstly, the topological variations are difficult to capture using a classical fully connected neural network (FCN) model [18]. When the grid topology changes, the trained model may lead to poor application performance in different grid topologies. Even though some algorithms employ graph neural networks, they cannot effectively leverage the physical knowledge of power systems. Secondly, model training requires a significant amount of time, and reinforcement learning encounters the sparse reward problem, leading to slow learning by the agent or even inability to learn the optimal strategy.

With consideration of the above problems, a graph attention network based PPO (GAT-PPO) algorithm for power systems with high proportion of wind power is proposed in this paper. Firstly, wind turbine nodes and load nodes with more active power are given larger weights. Then, the calculation of the attention coefficient in the GAT is improved based on the weights. Secondly, the improved ReliefF algorithm (defined as ReliefF-S algorithm hereafter) extracts the critical features affecting system stability. Then, these features are used as the state variables of the GAT-PPO algorithm to reduce the dimension of state space. Finally, the samples retained in the experience buffer are improved to mitigate the sparse reward problem. The optimization and control of voltage can be realized based on the above improvement strategy. The major contributions of this paper can be summarized as follows.

1) The proposed GAT-PPO algorithm integrates the physical information of the power system into the GAT, and it can give more attention to the important nodes during voltage regulation. Moreover, compared with the traditional DRL algorithms, the proposed GAT-PPO algorithm exhibits better transfer learning performance and greater adaptability to various grid topologies by integrating with GAT.

2) The critical state variables for DRL training are screened out based on the ReliefF-S algorithm, which can reduce the dimension of state space and improve the training efficiency of the algorithm. Additionally, in light of the practical issues of VVC, a reward function construction method based on the constraints first and objective later is proposed, providing a reference for related research.

3) Training samples retained in the experience buffer are improved to mitigate the sparse reward problem. In the improved training samples, some samples with large temporal difference errors are retained as positive experiences to guide the agent in training toward the correct direction. Meanwhile, a small number of samples that violate constraints are also retained, which act as negative experiences to warn the agent to avoid adopting the action strategies that violate constraints. An expanded boundary in terms of manageable grid topologies of the proposed GAT-PPO algorithm is found.

The rest of this paper is organized as follows. Section II introduces the DRL model for intraday VVC. The physics-informed GAT is presented in Section III. Section IV elaborates the state space reduction based on the ReliefF-S algorithm. Section V introduces intraday VVC based on GAT-PPO algorithm. Comparative studies are shown and discussed in Section VI. Finally, conclusions are drawn in Section VII.

II. DRL Model for Intraday VVC

The PPO algorithm is an improvement upon the trust region policy optimization (TRPO) algorithm, which is capable of handling continuous and discrete action spaces with good convergence [19]. The PPO algorithm consists of two neural networks: policy network and value network.

In the paper, the state space, action space, and reward function of the DRL model for VVC are designed and defined. The concise schematic diagram of the proposed GAT-PPO algorithm is shown in Fig. 1. The detailed settings are as follows.

A. Agent and Environment

The system operator or control program is set as an agent. The agent contains two neural networks: an improved GAT and an FCN. The power system dynamic simulator interacting with the agent is set as environment.

B. State Space

The grid state information in the model includes the wind turbine output, traditional unit output, load demand, voltage distribution, branch power distribution, reactive power output of dynamic reactive power compensation device, and adjacency matrix. The state space S_t is as follows:

$$S_t=\{P_t^W,Q_t^W,P_t^G,Q_t^G,P_t^L,Q_t^L,U_t,S_t^B,Q_t^D,\boldsymbol{Z}_t\} \tag{1}$$

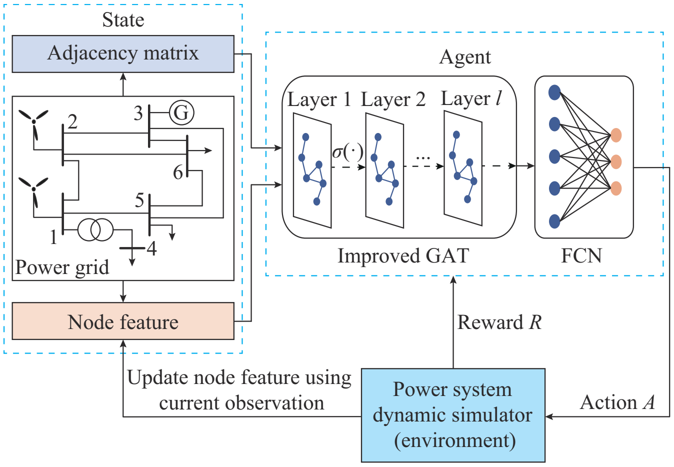


Fig. 1. Concise schematic diagram of proposed GAT-PPO algorithm.

where P_t^W and Q_t^W are the sets of active and reactive power outputs of all wind turbines at time t, respectively; P_t^G and Q_t^G are the sets of active and reactive power outputs of traditional units at time t, respectively; P_t^L and Q_t^L are the sets of active and reactive power demands of all loads at time t, respectively; U_t is the set of voltage amplitudes of all nodes at time t; S_t^B is the set of apparent power amplitudes of all branches at time t; Q_t^D is the set of reactive power outputs of all dynamic reactive power compensation devices at time t; and $\boldsymbol{Z}_t$ is the adjacency matrix of the system at time t.

C. Action Space

The action space represents the solution space. In this paper, the following VVC measures are considered: reactive power regulation of each static var generator (SVG) configured in the wind farm and reactive power regulation of each static var compensator (SVC). The action space A_t is defined as:

$$A_t=\{\Delta Q_t^S, \Delta Q_t^C\} \tag{2}$$

where ΔQ_t^S is the set of reactive power variations of all SVGs at time t; and ΔQ_t^C is the set of reactive power variations of all SVCs at time t.

D. Reward Function

The reward is crucial in guiding the learning direction of the agent. In the study, a reward function construction method based on the constraints first and objective later is designed in the paper. The main idea is to divide the reward function according to the constraint conditions and the objective function. The objective function is considered only when the states satisfy the constraints. Otherwise, a penalty value is returned.

In this paper, the reward is designed based on the network loss, and the network loss is the objective function of the model. The smaller the network loss, the larger the reward. The mathematical expression is:

$$R_{r,t}=\left(\frac{P_{loss,t}^{ref}-P_{loss,t}}{S_B}+e_r\right)\times 100\% \tag{3}$$

where $P_{loss,t}^{ref}$ is the preset network loss at time t; $P_{loss,t}$ is the actual network loss at time t; S_B is the base capacity, typically 100 MVA; and $e_r=0.25$ is the reward adjustment coefficient.

In this paper, the penalty is designed based on system security and stability constraints. The constraints consist of four parts: branch power constraints, node voltage constraints, generator reactive power constraints, and generator active power constraints. To ensure the safe and stable operation of the power system, the values of these variables need to be within specified ranges. When designing the penalty, the more serious the violation, the larger the penalty. Inspired by [20], the mathematical expression of the designed penalty is given as:

$$R_{f,t}=\lambda_{PB}\sum_{i=1}^{b}\left(\frac{S_{Bi,t}-S_{Bi}^{\max}}{S_{Bi}^{\max}}\right)+\lambda_U\sum_{i=1}^{n}\left|\frac{2U_{i,t}-U_i^{\max}-U_i^{\min}}{U_i^{\max}-U_i^{\min}}\right|+\lambda_P\sum_{i=1}^{k}\left|\frac{2P_{Gi,t}-P_{Gi}^{\max}-P_{Gi}^{\min}}{P_{Gi}^{\max}-P_{Gi}^{\min}}\right|+\lambda_Q\sum_{i=1}^{k}\left|\frac{2Q_{Gi,t}-Q_{Gi}^{\max}-Q_{Gi}^{\min}}{Q_{Gi}^{\max}-Q_{Gi}^{\min}}\right| \tag{4}$$

where λ_{PB}, λ_U, λ_P, and λ_Q are the penalty factors of branch power violation, node voltage violation, unit active power violation, and unit reactive power violation, respectively; b is the number of branches; n is the number of nodes; k is the number of generators; $S_{Bi,t}$ is the apparent power of branch i at time t; $S_{Bi}^{\max}$ is the maximum apparent power allowed by branch i; $U_{i,t}$ is the voltage of node i at time t; $U_i^{\max}$ and $U_i^{\min}$ are the maximum and minimum voltages of node i, respectively; $P_{Gi,t}$ is the active power of generator i at time t; $P_{Gi}^{\max}$ and $P_{Gi}^{\min}$ are the maximum and minimum active power of generator i, respectively; $Q_{Gi,t}$ is the reactive power of generator i at time t; and $Q_{Gi}^{\max}$ and $Q_{Gi}^{\min}$ are the maximum and minimum reactive power of generator i, respectively. Each item in (4) is normalized to eliminate the inconsistency problem of dimension and magnitude, the values of which range from 0 to 1. Through multiple tests with different cases, the values of all four penalty factors are set to be 0.25 in the paper.

According to the designed reward and penalty, the mathematical expression of the total reward function is given as:

$$R_t=\begin{cases}-R_{f,t} & s_t\notin S_{cons}\\ R_{r,t}-R_{f,t} & s_t\in S_{cons}\end{cases} \tag{5}$$

where S_{cons} is the set of state constraints; $R_{f,t}$ is the penalty given for violation of constraints, and its purpose is to enable the agent to make decisions within constraints; and $R_{r,t}$ is the reward given when all constraints are satisfied, enabling the agent to find the optimal decision based on the feasible decisions.

According to the mathematical expression of the total reward function, it can be observed that when any constraint is violated, the agent gets a negative reward according to (4). The purpose of this setting is to guide the agent to give an action that satisfies all constraints. When all constraints are not violated, it can be observed from (4) and (5) that the closer the node voltage is to 1 p.u., the larger the positive reward value the agent receives under the same other constraints. The purpose of this setting is to guide the agent to learn an action that can obtain an ideal voltage distribution.

III. Physics-informed GAT

A. GAT

The GAT introduces the attention mechanism into the graph neural network, which obtains the overall information of the network from the local information by calculating the importance of adjacent nodes to the central node. The advantage of the GAT is that it does not require any kind of costly matrix operation or depends on knowing the graph structure upfront, making it directly applicable to inductive learning issues [21]. Therefore, GAT has a strong transfer learning performance, which is conducive to being applied to voltage control in various grid topologies.

The expression of the attention coefficient is given as [21]:

$$\alpha_{ij} = \text{soft}\max_{j}\,(e_{ij}) = \frac{\exp(LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}h_i \| \boldsymbol{W}h_j]))}{\sum_{k\in M_i}\exp(LeakyReLU(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}h_i \| \boldsymbol{W}h_k]))} \tag{6}$$

where e_{ij} is the importance of node j to node i; $\boldsymbol{W}$ is the weight matrix; $\boldsymbol{a}$ is the parameter of a single-layer feedforward neural network; *LeakyReLU* is the activation function; M_i is the set of neighbor nodes of node $i\,(j\in M_i)$; h_i, h_j, and h_k are the characteristics of nodes i, j, and k, respectively; and $\|$ represents the concatenation operation.

B. Improvement of GAT

Extensive physical knowledge has been developed in power systems, and the application of the GAT in the field of electric power should be combined with the actual situation in the field. In power systems, the wind turbine nodes and load nodes are the key ones that affect voltage safety and stability in the power system with the integration of wind power. The higher the active power of these nodes, the more likely voltage safety and stability issues are to occur [22]. Therefore, the larger the active power of these nodes, the more attention should be paid during voltage regulation. In the paper, the attention to these nodes in the voltage regulation process is enhanced by assigning them larger weights in the attention coefficients. The improved GAT is called the physics-informed GAT.

The weight coefficients of wind turbine and load nodes are defined as:

$$\begin{cases} \beta^{W}_{i,t} = \dfrac{P^{W}_{i,t}}{P^{W}_{\min,t}} & P^{W}_{i,t}\neq 0 \text{ or } P^{L}_{i,t}\neq 0 \\ \beta^{L}_{i,t} = \dfrac{P^{L}_{i,t}}{P^{L}_{\min,t}} & P^{W}_{i,t}\neq 0 \text{ or } P^{L}_{i,t}\neq 0 \\ \beta^{W}_{i,t} = \beta^{L}_{i,t} = 0.9 & P^{W}_{i,t} = P^{L}_{i,t} = 0 \end{cases} \tag{7}$$

where superscripts W and L denote the sets of wind turbine nodes and load nodes, respectively; $\beta^{W}_{i,t}$ and $\beta^{L}_{i,t}$ are the weight coefficients of wind turbine node i and load node i at time t, respectively; $P^{W}_{i,t}$ and $P^{L}_{i,t}$ are the active power of wind turbine node i and load node i at time t, respectively; and $P^{W}_{\min,t}$ and $P^{L}_{\min,t}$ are the minimum active power of all wind turbines and the minimum active power of all loads at time t, respectively.

According to the above calculation method of weight coefficient, it can be observed that the larger the active power of node i, the larger its weight coefficient value. For each load or wind turbine node whose active power is not equal to 0, the weight coefficient is multiplied by the corresponding attention coefficient e_{ij}. For other nodes, e_{ij} is multiplied by 0.9. Therefore, the key wind turbine and load nodes can be paid more attention through the improved GAT.

IV. State Space Reduction Based on ReliefF-S Algorithm

One of the important causes for slow training and non-convergence of reinforcement learning is that the dimension of its state space or action space is too large. To solve the problem, a state space reduction strategy is proposed. The key features that have a great influence on voltage stability are obtained through the method of key feature extraction, and then these key features are taken as state variables of reinforcement learning. The method can reduce the state space dimension and accelerate the convergence of the model.

The ReliefF algorithm is an efficient feature extraction algorithm that assigns weights to features based on their correlation with the labels. The feature whose weight is less than the setting threshold value will be removed, and then the optimal feature subset can be obtained [23]. Since the algorithm does not consider the correlation among samples when determining the homogeneous and heterogeneous samples, it may lead to inaccurate judgment of the homogeneous and heterogeneous samples, ultimately affecting the identification of key features. Furthermore, the correlation between two samples is not considered when calculating the weight.

To address the above problems, the Spearman correlation coefficient is adopted to improve the ReliefF algorithm as the ReliefF-S algorithm. The reason for using the Spearman correlation coefficient is that it does not require a specific distribution between variables. The specific methods of the improvement are described as follows.

Firstly, the Spearman correlation coefficient ρ [24], as shown in (8), is used to replace the original sample category judgment formula of the ReliefF algorithm, thereby helping accurately identify homogeneous and heterogeneous samples. Secondly, (8) is multiplied by the sample distance $d(\cdot)$ [23], so the overall correlation among samples is considered when calculating the weight contribution. The expression of improved weight is shown in (9).

$$\rho = \frac{\sum_{i=1}^{n_\rho}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n_\rho}(x_i-\bar{x})^2\,(y_i-\bar{y})^2}} \tag{8}$$

$$w'_{f_i} = w'_{f_i} - \sum_{j=1}^{g}\frac{\rho d(x_i,H_j,f_i)}{mk} + \sum_{l\notin l_{x_i}}\sum_{j=1}^{g}\frac{P(l)}{1-P(l_{x_i})}\frac{\rho d(x_i,M_j,f_i)}{mg} \tag{9}$$

where x_i and y_i represent two samples, respectively; $\bar{x}$ and $\bar{y}$ represent the average values of the two samples, respectively; n_ρ is the total number of features in the sample; f_i is the feature; w'_{f_i} is the weight of f_i; H_j and M_j are the jth near-neighbor homogeneous samples and heterogeneous samples of sample x_i, respectively; $d(x_i,H_j,f_i)$ or $d(x_i,M_j,f_i)$ represents the distance between samples x_i and H_j or M_j on f_i, respectively; m is the number of iterations of the algorithm; g is the number of near-neighbor homogeneous samples; $P(l)$ is the probability of label l; l_{x_i} is the label of sample x_i; $P(l_{x_i})$ is the probability that sample x_i belongs to some kind of label; $\sum_{j=1}^{g}\frac{d(x_i,H_j,f_i)}{mk}$ is the weight contribution of sample x_i and g near-neighbor homogeneous samples on f_i; and $\sum_{l\notin l_{x_i}}\sum_{j=1}^{g}\frac{P(l)}{1-P(l_{x_i})}\frac{d(x_i,M_j,f_i)}{mg}$ is the weight contribution of sample x_i and all near-neighbor heterogeneous samples on f_i.

Finally, the key factors affecting the system voltage stability are screened out based on the proposed ReliefF-S algorithm. These key factors are the state variables used in DRL training.

V. Intraday VVC Based on GAT-PPO Algorithm

A. Design of Experience Buffer

A lack of effective reward information will lead to slow learning or even failure to learn the optimal strategy. To mitigate the sparse reward problem, the samples retained in the experience buffer are improved in this paper. The designed experience buffer retains both samples with a large temporal difference error and a small number of samples that violate voltage constraints. The former is used as a positive experience to guide the agent to train in the right direction, while the latter is used as a negative experience to warn the agent to avoid action strategies that violate constraints.

B. GAT-PPO Algorithm for Intraday VVC

This subsection introduces the structure and the training process of the proposed GAT-PPO algorithm.

The structure of the proposed GAT-PPO algorithm is shown in Fig. 2. It mainly includes three parts: the improved GAT, policy network, and value network. Unlike the traditional PPO algorithm, the state information of the proposed GAT-PPO algorithm needs to output node features through the GAT. The GAT contains l layers. The policy network and value network are constructed by the deep neural network, and their detailed structures are as follows. The policy network is responsible for the sequential decisions of VVC, which consists of two parts: the policy layer and the action layer. During training, the observed power system state variables are first input into the GAT to generate the node feature set. Then, the node feature set is input into the policy layer for training. Finally, the output of the policy layer is input to the action layer, and the action layer outputs action.

The value network maps the system state S_t to the expected future cumulative rewards, which contains a state value layer. During training, the observed power system state variables are first input into the GAT to generate the node feature set. Then, the node feature set is input into the state value layer for training. Finally, the state value layer outputs the state value function.

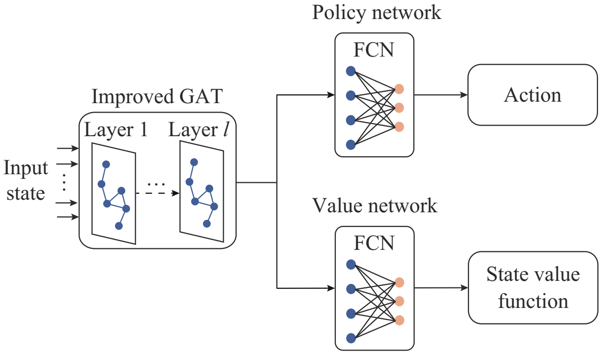


Fig. 2. Structure of proposed GAT-PPO algorithm.

The objective function of conventional policy gradient based DRL optimization is given as [19]:

$$L^P(\theta)=\hat{E}[\log_2 \pi_\theta(a_t|s_t)\hat{A}_t] \tag{10}$$

where θ is the policy parameter; $\hat{E}[\cdot]$ represents the empirical average over finite samples; s_t and a_t are the state and action at time t, respectively; π_θ is a stochastic policy; and $\hat{A}_t$ is an estimator of the advantage function at time t.

The objective function of the value function $L^V(\cdot)$ can be formulated as:

$$L^V(\phi)=\hat{E}[\hat{V}_\phi^{target}(s_t)-V_\phi(s_t)] \tag{11}$$

$$\hat{V}_\phi^{target}(s_t)=r_{t+1}+\gamma V_\phi(s_{t+1}) \tag{12}$$

where ϕ is the value function parameter; $\gamma\in[0,1]$ is the discount factor; $V_\phi(s_t)$ is the state value function at time s_t; r_{t+1} is the reward at time $t+1$; and $\hat{V}_\phi^{target}(\cdot)$ is the target value of TD error, and the parameter can be updated by the stochastic gradient descent algorithm according to the gradient $\nabla L^V(\phi)$.

The training process of the proposed GAT-PPO algorithm is shown in Fig. 3, where $r_t(\theta)=\pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$ denotes the ratio of probability of action a_t under the new policy and old policy; and $L^{CLIP}(\theta)$ indicates that the clipping mechanism in [19] is used to constrain the variation range of $r_t(\theta)$. The training of the improved GAT is an end-to-end process. The improved GAT and the FCN are trained at the same time, and the error is transmitted and optimized through the backpropagation in the entire model. During training, the policy network continuously interacts with the VVC training environment, and the environment sends the experience tuples $\langle s_t,a_t,r_{t+1},s_{t+1}\rangle$ to the experience buffer to form finite samples. Then, the samples are transmitted to the policy network and the value network. The update processes of the policy network and the value network are described as follows.

1) Policy Network

The sequences of power system state variables are separately input into two action networks, resulting in two policy distributions. Among them, one is for the new policy and the other is for the old policy. According to the new and old policy distributions, the probabilities for selecting each action under both policies are computed separately. Then, the probability of the new policy is divided by the probability of the old policy to obtain the ratio of policy probability $r_t(\theta)$. The objective function value of the improved GAT-PPO algorithm is calculated using the advantage function and the ratio of policy probability. Subsequently, the negative of the objective function is taken as the loss function of the neural network. The parameters of the policy network are updated through backpropagation using the loss function. Finally, a new policy satisfying the clipping requirements is obtained.

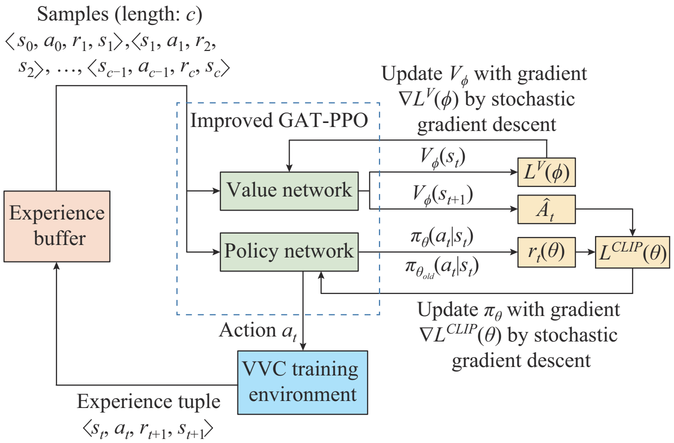

Samples (length: c)
$\langle s_0, a_0, r_1, s_1\rangle, \langle s_1, a_1, r_2, s_2\rangle, \ldots, \langle s_{c-1}, a_{c-1}, r_c, s_c\rangle$
Update V_ϕ with gradient $\nabla L^V(\phi)$ by stochastic gradient descent
Improved GAT-PPO
$V_\phi(s_t)$
$L^V(\phi)$
Experience buffer
Value network
$V_\phi(s_{t+1})$
$\hat{A}_t$
Policy network
$\pi_\theta(a_t|s_t)$
$r_t(\theta)$
$L^{CLIP}(\theta)$
$\pi_{\theta_{old}}(a_t|s_t)$
Update π_θ with gradient $\nabla L^{CLIP}(\theta)$ by stochastic gradient descent
Action a_t
VVC training environment
Experience tuple
$\langle s_t, a_t, r_{t+1}, s_{t+1}\rangle$

Fig. 3. Training process of proposed GAT-PPO algorithm.

2) Value Network

The observed power system state variables are input into the value network to obtain the corresponding value function. The discounted reward is computed based on the discount reward calculation formula $G_t = \sum \gamma^u r_{t+u+1}$ ($u = 0, 1, \ldots, \infty$). Then, the advantage function is calculated. The value network parameters are updated by computing gradient with respect to the function in (11) and performing backward propagation.

When the loss function values of both the policy network and the value network are stable and close to a small value, and the moving average reward is positive and tends to stabilize, it indicates that the algorithm has converged.

VI. Case Studies

A. Case Overview

In the paper, the modified IEEE 39-bus system and an actual power grid are analyzed as cases. The single-line diagram of the modified IEEE 39-bus system is shown in Fig. 4, and the symbol "W" and "G" represent the wind turbine and traditional generator, respectively. The modifications of the IEEE 39-bus system are described as follows. The traditional generators at buses 32, 35, 37, 38, and 39 are replaced by wind turbines. After replacement, the penetration rate of wind power is 70%, which means that the system is a power system with high proportion of wind power. The SVGs are configured at each wind turbine bus with a capacity range of ±20% of the total active power of wind turbines at each bus. The SVCs are configured at buses 2, 3, 9, 12, 25, 27, and 29 with a capacity range of ±150 Mvar. Loads on buses 31 and 39 are removed. The normal voltage ranges of the load bus and the traditional generator bus are set to be 0.95-1.05 p.u. and 0.9-1.1 p.u., respectively. Considering the situation that the wind turbine bus should have a certain voltage stability margin, its normal voltage range is set to be 0.99-1.06 p.u.. The actual power grid is a power system with the integration of wind power comprising 222 bus nodes. The input dimensions of the GAT are 39×6 and 222×6 in the modified IEEE 39-bus system and the actual power grid, respectively. The hidden layer dimension is 8, and the activation function is *LeakyReLU*. The number of heads in the GAT is 8. The output dimensions of the GAT are 39×2 and 222×2 in the modified IEEE 39-bus system and the actual power grid, respectively. The state space dimensions of the two systems are 220 and 1102, respectively. The action space dimensions of the two systems are 12 and 40, respectively.

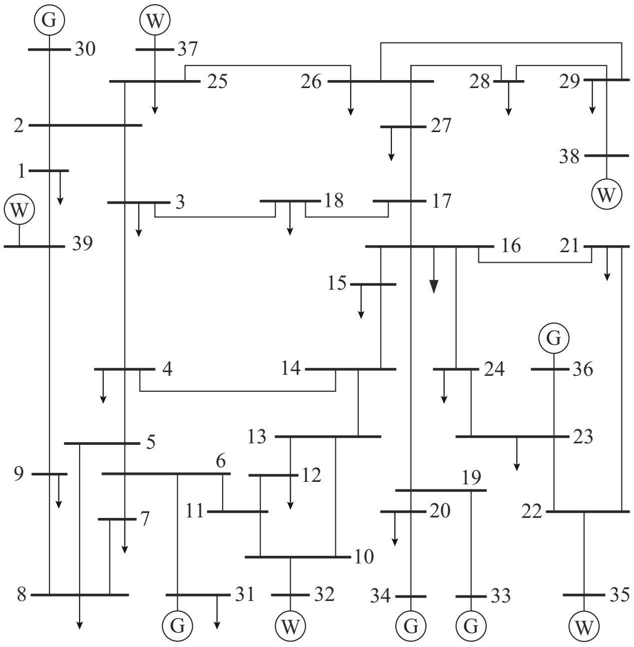


Fig. 4. Single-line diagram of modified IEEE 39-bus system.

The simulation environment of power systems is provided by PSSE software in the paper. The annual operation data of an actual power system with the integration of wind power are scaled to the modified IEEE 39-bus system, generating numerous operation data. Meanwhile, the forecasting data of load and wind power are processed as follows. The sample data are generated by setting the forecasting errors from the operation data of the modified IEEE 39-bus system, with a maximum forecasting error of 20% for wind power and 15% for load power. To change the operation conditions during algorithm training, different grid topologies are selected in the two systems. The grid topology changes are achieved by disconnecting the following transmission lines one at a time. In the modified IEEE 39-bus system, four different grid topologies are selected: ① no disconnecting; ② the line between bus 5 and bus 6; ③ the line between bus 16 and bus 24; and ④ the line between bus 22 and bus 23. In the actual power grid, eight different grid topologies are randomly selected in the same way. Then, the DRL algorithm is trained based on these data.

B. Screening of State Variables Based on ReliefF-S Algorithm

The modified IEEE 39-bus system is used as a case to illustrate the screening of state variables. To obtain the sample data, the active power margin of the system is calculated by using the annual operation data in the modified IEEE 39-bus system. When the active power margin of the sample is larger than 10%, the sample is a voltage-stable sample. Otherwise, the sample is a voltage-unstable sample. In this paper, 16529 voltage-stable samples and 15951 voltage-unstable samples are obtained. Since the branch power is allowed to exceed the limit to a certain extent without affecting the voltage stability, the branch power can be screened. The weights of 34 branches are shown in Table I.

TABLE I
WEIGHTS OF 34 BRANCHES

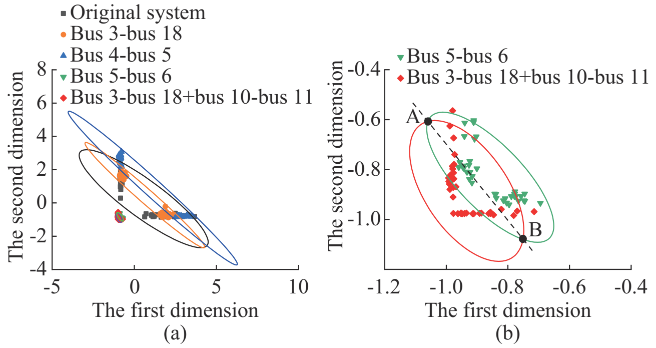
From bus	To bus	Weight	From bus	To bus	Weight
4	5	0.2176	1	2	0.1236
10	13	0.2073	2	25	0.1058
13	14	0.2048	4	14	0.1057
5	6	0.2047	23	24	0.1050
6	7	0.2008	8	9	0.1038
6	11	0.1902	9	39	0.1038
25	26	0.1840	26	29	0.1014
5	8	0.1764	26	28	0.0970
10	11	0.1762	2	3	0.0918
3	4	0.1687	26	27	0.0880
14	15	0.1653	3	18	0.0857
1	39	0.1434	17	18	0.0824
16	21	0.1420	17	27	0.0792
15	16	0.1418	28	29	0.0694
21	22	0.1373	16	19	0.0625
16	24	0.1284	22	23	0.0576
7	8	0.1245	16	17	0.0419

It can be observed from Table I that the weight values of the 34 branches vary greatly, with the maximum value being 5.2 times the minimum value. It indicates that different branches contribute differently to voltage stability. The average weight of 34 branches is 0.1299. Meanwhile, it is noticeable that the weight values are mostly concentrated above 0.1. Therefore, 0.1014 is selected as the weight threshold. Finally, the apparent power of 24 branches is retained as the state variables based on the weight threshold. Through the above processing, the state space dimension of DRL can be reduced by ten dimensions.

C. GAT Output Under Grid Topology Change

To illustrate the impact of grid topology changes on DRL training, meanwhile, in view of the issues studied in the paper, to identify the topological boundary that the GAT-PPO algorithm can handle, the output features of GAT in the modified IEEE 39-bus system are extracted for comparison.

In the paper, the grid topology is changed by disconnecting branches, and the GAT output under different grid topologies is shown in the matrix scatterplot in Fig. 5. An example is shown to illustrate the new grid topologies formed by disconnecting branches of bus 3-bus 18, bus 4-bus 5, bus 5-bus 6, as well as simultaneously disconnecting branches of bus 3-bus 18 and bus 10-bus 11. Among them, Fig. 5(b) is a locally enlarged graph of the two graphs in the lower left corner of Fig. 5(a). Note that the confidence ellipse formed by disconnecting branch of bus 5-bus 6 is the farthest from the confidence ellipse formed by the original system. The grid topology formed by simultaneously disconnecting branches of bus 3-bus 18 and bus 10-bus 11 is a topology that the proposed GAT-PPO algorithm cannot handle. Meanwhile, the confidence ellipse formed by the topology is closest in distance to the confidence ellipse formed by disconnecting branch of bus 5-bus 6.



(a)

(b)

Fig. 5. Matrix scatterplots under different grid topologies. (a) Original graph. (b) Local enlarged graph.

According to Fig. 5(a), the following two conclusions can be drawn: ① the GAT output changes after the grid topology is changed; and ② after the grid topology is changed, if the numerical values of the features output by GAT are on the left side of the numerical values of the original system features, there may be a grid topology that the proposed GAT-PPO algorithm cannot handle.

According to Fig. 5(b), it can be observed that the intersection points of the two topologies are A and B with the corresponding two-dimensional coordinates of [–1.0578, –0.6063] and [–0.7499, –1.0730], respectively. In a certain grid topology, if the numerical values of GAT features are primarily concentrated to the left of the line AB, the proposed GAT-PPO algorithm will not be able to handle the situation.

D. Comparative Analysis of Algorithm Performance in Modified IEEE 39-bus System

To verify the effectiveness of the proposed GAT-PPO algorithm, the comparative analyses are carried out from the perspectives of different algorithms and different voltage scenarios. In terms of algorithm comparison, the proposed GAT-PPO algorithm is compared with the PPO algorithm, TD3 algorithm, particle swarm optimization (PSO)-based SP algorithm, and genetic algorithm (GA)-based SP algorithm. Regarding voltage scenarios, the branch of bus 3-bus 18 is disconnected in the paper, which indicates that the grid topology is changed. Then, two scenarios of high voltage and low voltage are selected for comparison. In these two scenarios, the prediction data of wind power output and load demand are randomly generated according to their fluctuation range. The VVC based on the prediction data is applied to the corresponding actual scenario without random processing, thereby comparing the performance of the proposed algorithm under the uncertainties of wind power output and load demand. The comparative analyses include the following four aspects: training speed, importance features of nodes, VVC performance, and control performance under different grid topologies.

1) Training Speed

Training speed is an important index to measure the superiority of the proposed GAT-PPO algorithm. The faster the training speed, the more conducive to the online application of the proposed GAT-PPO algorithm. In the paper, the proposed GAT-PPO algorithm is compared with PPO algorithm and TD3 algorithm. Meanwhile, to verify the effectiveness of the dimension reduction strategy, the proposed GAT-PPO algorithm is compared with the GAT-PPO algorithm without dimension reduction (referred to as "GAT-PPO-wdr").

The information entropy is used to measure the training speed of four kinds of DRL algorithms [25]. The information entropy curves of each algorithm are shown in Fig. 6(a). In addition, the reward curves of four kinds of DRL algorithms during training are shown in Fig. 6(b), where the reward function of four DRL algorithms adopts the construction method proposed in the paper.

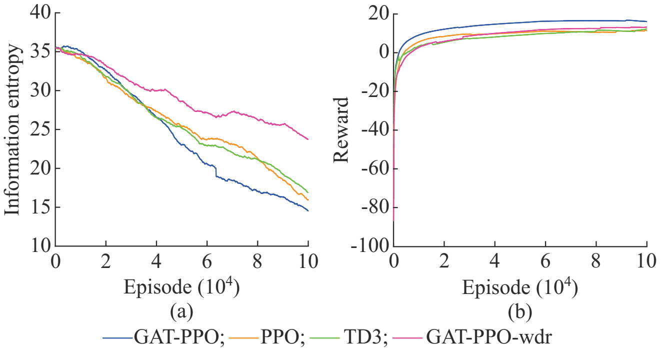


Fig. 6. Information entropy and reward curves of each algorithm in modified IEEE 39-bus system. (a) Information entropy curves. (b) Reward curves.

As depicted in Fig. 6(a), the information entropy curve of the proposed GAT-PPO algorithm drops the fastest, followed by the PPO algorithm and TD3 algorithm, and the GAT-PPO-wdr algorithm drops the slowest. Meanwhile, the proposed GAT-PPO algorithm achieves the lowest entropy value at the end of training. The above results show that: ① the training speed of the proposed GAT-PPO algorithm is higher than those of the other three DRL algorithms; and ② the training speed of the GAT-PPO-wdr algorithm is significantly lower than those of the other three DRL algorithms with dimension reduction. The reasons for these phenomena are as follows: ① the proposed GAT-PPO algorithm improves the experience buffer, which can guide the agent to accelerate the training; and ② reducing the state space dimension can effectively improve the training efficiency. Therefore, the training speed and convergence performance of the proposed GAT-PPO algorithm are better than those of other three DRL algorithms, and it has better online application value.

According to Fig. 6(b), the reward value of the proposed GAT-PPO algorithm is larger than those of other algorithms. It shows that the proposed GAT-PPO algorithm can better guide the agent training. Meanwhile, the reward curve of the GAT-PPO-wdr algorithm rises at the slowest rate. The phenomenon also indicates that when the dimensionality of the state space is not reduced, and the training speed of the algorithm is affected. In addition, during the training process, both the PPO algorithm and the TD3 algorithm exhibit situations where their reward values surpass those of the other, which indicates that both algorithms can achieve better control results than the other at different time. However, regarding the issue studied in the paper, the final reward values of the two algorithms do not differ significantly.

2) Importance Features of Nodes

Combined with the VVC problem in the paper, the nodes with larger active power in the wind turbine nodes and load nodes are more important. These nodes should be given higher priority during voltage regulation. Meanwhile, the higher-priority nodes also represent the primary nodes in the graph. To compare the importance features of the nodes in the graph, the low-voltage scenario is taken as a case, and the GCN-based PPO (GCN-PPO) algorithm is added for comparison. The first six load nodes with larger active power and the first three wind turbine nodes are selected for analysis in the adopted case. The comparison of node voltages under different algorithms is shown in Table II, where the load nodes and the wind turbine nodes are arranged in descending order according to the active power.

TABLE II
COMPARISON OF NODE VOLTAGES UNDER DIFFERENT ALGORITHMS

Node type	Bus	Active power (MW)	Voltage (p.u.)			
			GAT-PPO	PPO	TD3	GCN-PPO
Load node	7	833.8	1.0092	1.0181	1.0336	1.0163
	8	822.0	1.0038	1.0184	1.0309	1.0132
	20	680.0	1.0012	0.9982	0.9792	0.9901
	4	600.0	1.0068	1.0166	1.0103	1.0120
	16	329.0	0.9965	0.9893	0.9879	0.9897
	3	322.0	1.0006	1.0319	1.0028	1.0169
Wind turbine node	37	1395.0	1.0208	1.0329	1.0439	1.0392
	39	1000.0	1.0250	1.0438	1.0326	1.0403
	38	830.0	1.0019	1.0110	1.0452	1.0104

According to Table II, compared with other algorithms, the proposed GAT-PPO algorithm can make the voltages of the primary nodes closer to 1 p.u., which indicates better voltage distribution. The GCN-PPO algorithm does not ensure that the voltage of all primary nodes is closer to 1 p.u. than those of the PPO and TD3 algorithms. The reason for this phenomenon is that the proposed GAT-PPO algorithm integrates the physical knowledge of power systems in voltage regulation, which gives higher priority to the primary nodes during voltage regulation. Therefore, the voltage of primary nodes is prioritized to be restored to normal.

3) VVC Performance

The comparison of network loss for each algorithm is shown in Table III. The comparison diagrams of node voltages and voltage violation nodes for each algorithm are shown in Figs. 7 and 8, respectively. Moreover, the effectiveness of the continuous 6-hour VVC is validated using bus 4 as a case, as shown in Fig. 9. The time interval is 15 min. To observe whether the voltage has returned to the normal range, the boundary lines representing the upper and lower limits of the voltage according to the actual needs are marked. As shown by the dotted line in Figs. 7 and 8, the blue, orange, and green dotted lines represent 1.06 p. u., 1.05 p. u., and 0.99 p.u., respectively.

TABLE III
COMPARISON OF NETWORK LOSS FOR EACH ALGORITHM

Algorithm	Network loss (MW)	
	High-voltage case	Low-voltage case
Original system	78.127	170.689
GAT-PPO	72.702	152.264
PPO	75.539	157.883
TD3	76.650	156.288
PSO	75.155	157.327
GA	76.358	157.496

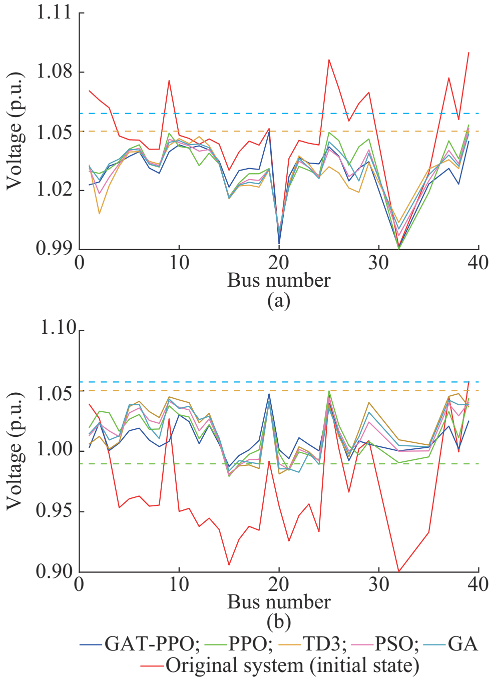


Fig. 7. Comparison of node voltages for each algorithm. (a) High-voltage case. (b) Low-voltage case.

It can be observed from Fig. 7(a) that each algorithm can restore the voltage of each node to the normal range. Meanwhile, for the node voltages optimized by each algorithm, there will be a situation where some node voltages obtained by one algorithm are closer to 1 p.u. than those obtained by other algorithms. However, in terms of the voltage recovery of the high-voltage nodes shown in Fig. 8(a), most of the node voltages obtained by the proposed GAT-PPO algorithm are much closer to 1 p.u. than those obtained by the PPO algorithm, indicating better voltage control performance. Meanwhile, although the TD3 algorithm obtains more nodes with voltage values close to 1 p.u. than the GAT-PPO algorithm, the proposed GAT-PPO algorithm achieves a 5.2% lower network loss than the TD3 algorithm. This indicates that the proposed GAT-PPO algorithm has better economy. Furthermore, the VVC performance based on the PSO algorithm and GA algorithm is similar, with their performance falling between the PPO algorithm and TD3 algorithm. However, they require excessively long computation time to obtain VVC strategy. For example, both PPO and TD3 algorithms take more than 150 s, while the proposed GAT-PPO algorithm takes less than 1 s. Additionally, it can be observed from Table III that the proposed GAT-PPO algorithm achieves the lowest network loss, which is 6.9% lower than that of the original system. Therefore, in terms of voltage control performance and economy, the overall VVC performance of the proposed GAT-PPO algorithm is superior under high-voltage condition.

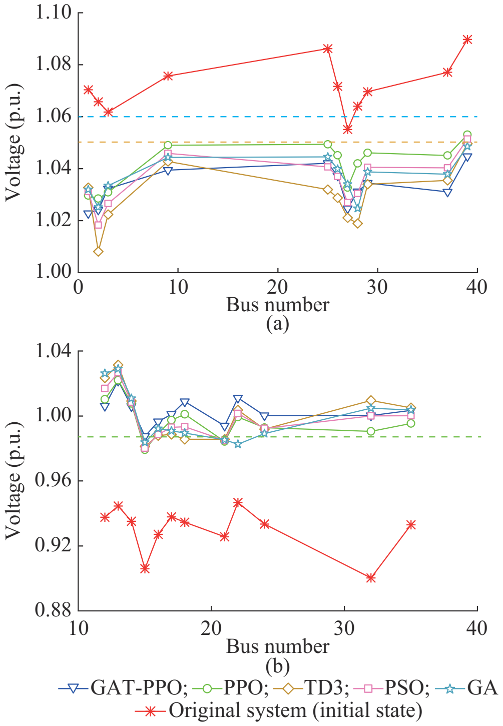


Fig. 8. Comparison of voltage violation nodes for each algorithm. (a) High-voltage case. (b) Low-voltage case.

It can be observed from Fig. 7(b) that each algorithm can restore the voltage of each node to the normal range. Meanwhile, the node voltage optimized by each algorithm also presents the same situation, as shown in Fig. 7(a). However, it can be observed from Fig. 8(b) that the proposed GAT-PPO algorithm obtains more nodes with voltage values close to 1 p. u. than the PPO algorithm and the TD3 algorithm, which indicates that the proposed GAT-PPO algorithm has better voltage control performance. Meanwhile, the VVC performance based on the PSO algorithm and the GA algorithm is inferior to that of the PPO algorithm, and they require excessively long computation time to obtain the VVC strategy.

The solution time based on the PSO algorithm and the GA algorithm also exceeds 150 s. Furthermore, it can be observed from Table III that the proposed GAT-PPO algorithm achieves the lowest network loss, which is 10.8% lower than the network loss of the original system. Compared with other algorithms, the proposed GAT-PPO algorithm has a range of 2.6% to 4% lower network loss, indicating its economic efficiency. Therefore, the overall VVC performance of the proposed GAT-PPO algorithm is superior under low-voltage condition.

It can be observed from Fig. 9 that each algorithm can restore the voltage to the normal range. Meanwhile, within the continuous time period, the voltage distribution of bus 4 obtained by the proposed GAT-PPO algorithm is more ideal, which indicates that the VVC performance of the proposed GAT-PPO algorithm is better.

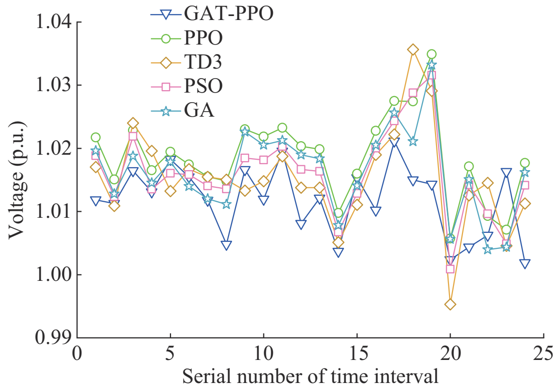


Fig. 9. Comparison of continuous 6-hour VVC.

Moreover, it can be observed that the proposed GAT-PPO algorithm can effectively cope with the uncertainty of the power system. The primary reason is that the agent has learned the patterns of changes in load demand and wind power output during the training process, and has mastered their probability distribution. Thus, the agent gives optimal control from the perspective of expectation.

4) Control Performance Under Different Grid Topologies

To further verify the transfer learning capability of the proposed GAT-PPO algorithm under different grid topologies, the remaining alternating current (AC) branches are disconnected in turn, resulting in a total of 30 new grid topologies. In each grid topology, two cases involving high voltage and low voltage are selected first, and then the VVC performance of the proposed GAT-PPO algorithm and the PPO algorithm is compared. Since the power flow will not converge when branches of bus 1-bus 39, bus 2-bus 3, bus 3-bus 4, bus 2-bus 25, bus 8-bus 9, bus 9-bus 39, bus 15-bus 16, bus 16-bus 19, and bus 28-bus 29 are disconnected, 21×2 test scenarios are generated finally. The comparison of network loss difference under different grid topologies is shown in Fig. 10. Note that the red column represents the case that the PPO algorithm cannot achieve VVC, and it does not represent the network loss difference.

According to Fig. 10, the proposed GAT-PPO algorithm can realize VVC under different new grid topologies, while the PPO algorithm fails to achieve VVC under four new grid topologies. The results prove that the proposed GAT-PPO algorithm has better transfer learning capability and adaptability to different grid topologies.

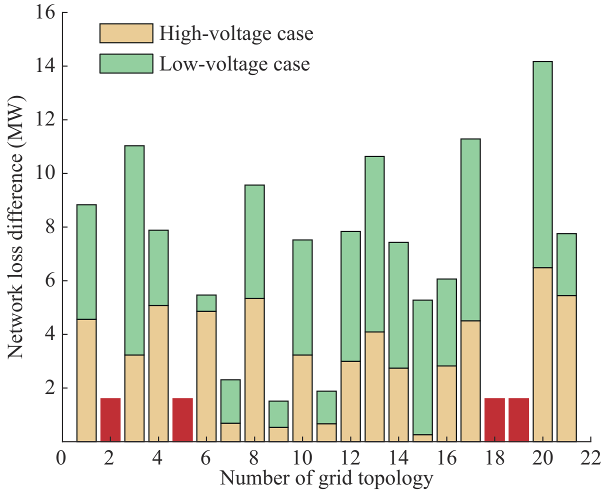


Fig. 10. Comparison of network loss difference under different grid topologies.

E. Comparative Analysis of Algorithm Performance in Actual Power Grid

An actual power grid is adopted to verify the effectiveness of the proposed GAT-PPO algorithm in the large scale system. The actual power grid contains 222 bus nodes and 285 AC branches. The analyses will be conducted from three aspects: training speed, VVC performance, and control performance under different grid topologies.

1) Training Speed

For the actual power grid, the ReliefF-S algorithm is adopted to remove a total of 61 branches. As a result, the dimension of the state space has been reduced by 61 dimensions. The information entropy and reward curves of each algorithm in actual power grid are shown in Fig. 11.

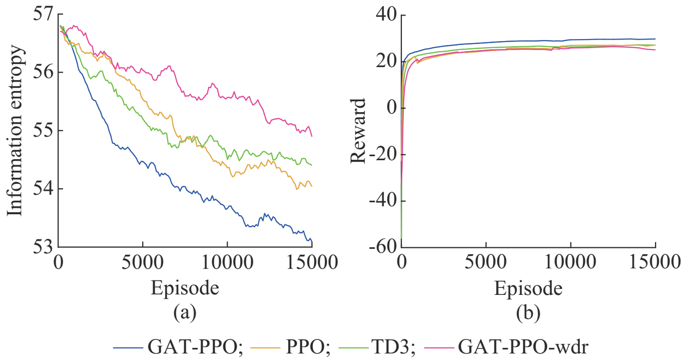


Fig. 11. Information entropy and reward curves of each algorithm in actual power grid. (a) Information entropy. (b) Reward.

It can be observed from Fig. 11(a) that the information entropy curve of the proposed GAT-PPO algorithm decreases the fastest, followed by the TD3 algorithm and the PPO algorithm, while the entropy curve of the GAT-PPO-wdr algorithm decreases the slowest. At the end of training, the proposed GAT-PPO algorithm has the smallest entropy value.

The reason is that the proposed GAT-PPO algorithm has improved the experience buffer, which can effectively guide the agent to accelerate training. Therefore, the training speed and convergence of the proposed GAT-PPO algorithm are superior to other DRL algorithms, offering better value for online applications.

According to Fig. 11(b), it can be observed that both the PPO algorithm and the TD3 algorithm have higher reward values during the training process. Therefore, the performance of the two algorithms in guiding agent training is similar. However, the reward value of the proposed GAT-PPO algorithm is higher than those of the PPO and TD3 algorithms, indicating that the proposed GAT-PPO algorithm can better guide agent training.

2) *VVC Performance*

A high-voltage case is used for comparison and analysis. In this case, one branch is disconnected, indicating the grid topology is changed. The comparison of node voltage obtained by each algorithm is shown in Fig. 12, and the comparison of network loss for each algorithm is presented in Table IV.

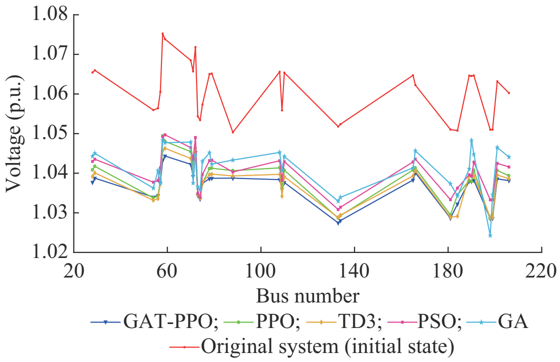


Fig. 12. Comparison of node voltage obtained by each algorithm.

TABLE IV
COMPARISON OF NETWORK LOSS FOR EACH ALGORITHM IN ACTUAL POWER GRID

Algorithm	Network loss (MW)
Original system	56.199
GAT-PPO	51.847
PPO	53.596
TD3	54.055
PSO	55.388
GA	54.163

According to Fig. 12, it can be observed that the proposed GAT-PPO algorithm obtains more nodes with voltage values close to 1 p.u. than the PPO algorithm and the TD3 algorithm, which indicates better voltage control performance of the proposed GAT-PPO algorithm. Furthermore, the voltage control performance based on the GA algorithm is the poorest.

According to Table IV, it can be observed that the proposed GAT-PPO algorithm achieves the lowest network loss, which is 7.7% lower than that of the original system. This indicates that the proposed GAT-PPO algorithm has better economic efficiency. Therefore, the comprehensive performance of the VVC of the proposed GAT-PPO algorithm is superior.

3) *Control Performance Under Different Grid Topologies*

To further validate the adaptability of the proposed GAT-PPO algorithm, the remaining AC branches are sequentially disconnected, and a total of 173 new grid topologies is eventually formed. Under each grid topology, high-voltage cases are selected, and then the VVC performance of the proposed GAT-PPO algorithm and the PPO algorithm is compared. The comparison of network loss difference under different grid topologies is shown in Fig. 13. Note that the blank spaces on the horizontal axis indicate the scenarios where the PPO algorithm cannot achieve VVC.

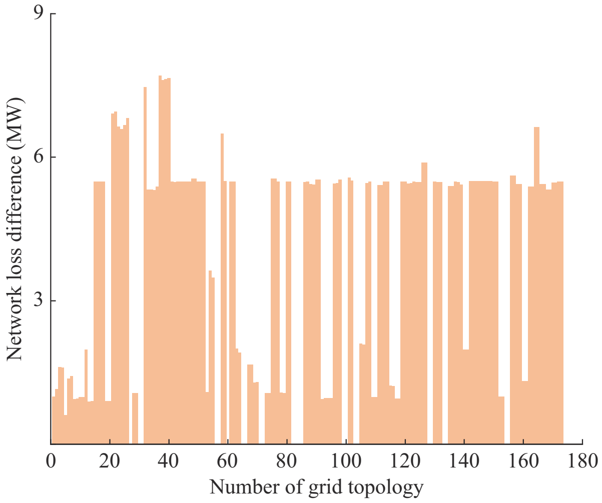


Fig. 13. Comparison of network loss difference under different grid topologies in actual power grid.

According to Fig. 13, the proposed GAT-PPO algorithm can achieve VVC under all new grid topologies, while the PPO algorithm fails to achieve VVC under 24 new grid topologies. Therefore, the proposed GAT-PPO algorithm demonstrates better adaptability to different grid topologies.

F. *Influence of DRL Algorithm Parameters on Result*

The performance of the DRL algorithm in this paper is affected by hyperparameters such as the discount factor γ and the number of neurons in the neural network. The value range of γ is usually 0.9-1. The larger the value of γ, the more the longer-term considerations of the agent, and the training difficulty of the proposed GAT-PPO algorithm also increases. The smaller the value of γ, the more the agent focuses on immediate gains, and the training difficulty of the proposed GAT-PPO algorithm decreases. Therefore, it is important to choose an appropriate γ when training the agent. The reward curves of different discount factors are shown in Fig. 14.

It can be observed from Fig. 14 that the final reward value is the largest when the discount factor $\gamma=0.95$, indicating the best training performance. The training performance is poorer when the $\gamma=0.92$, while a training failure occurs when $\gamma=0.98$. The reason is that the larger the value of γ, the harder it is to train. The discount factor $\gamma = 0.95$ is also that ultimately used in the paper after extensive testing.

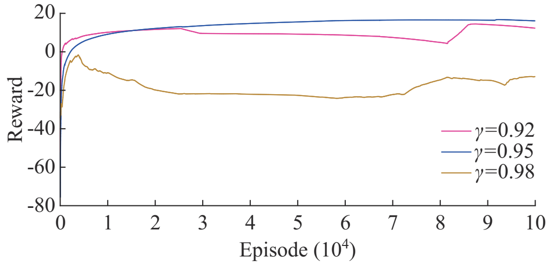


Fig. 14. Reward curves of different discount factors.

In addition, the paper sets the number of neurons in the neural network to be 64 based on extensive testing. The number of neurons is also crucial for algorithm training. If the number of neurons is too small, such as 16, it may prevent the neural network from learning correctly. Conversely, if the number of neurons is too large, such as 256, it may lead to an excessive number of parameters that the neural network needs to train, thus increasing the learning difficulty and affecting the network generalization ability.

VII. Conclusion and Future Work

This paper proposes GAT-based deep reinforcement learning for VVC of topologically variable power system, which incorporates voltage stability characteristics of system nodes into the attention mechanism, prioritizing essential nodes during voltage regulation. Furthermore, the challenges of slow training and sparse reward in DRL are effectively mitigated through the ReliefF-S algorithm and the optimization of the experience buffer, respectively.

According to the results of case studies, it can be observed that the proposed GAT-PPO algorithm not only has rapid convergence speed and good adaptability to different grid topologies but also possesses a strong ability to cope with uncertainties. The proposed GAT-PPO algorithm reduces the network loss by 6.9%, 10.8%, and 7.7%, respectively, demonstrating favorable economic outcomes. The proposed GAT-PPO algorithm can obtain an agent with strong transfer learning capability using a small amount of grid topology data, without the need for data from all different grid topologies. Meanwhile, the design of the reward function, prioritizing constraints first and objective later, closely aligns with practical VVC challenges. Additionally, the proposed GAT-PPO algorithm showcases an expanded boundary in terms of manageable grid topologies. In summary, the proposed GAT-PPO algorithm has better VVC performance, which strongly supports engineering applications.

The proposed GAT-PPO algorithm encounters two limitations: the curse of dimensionality when facing large-scale power grids, and the performance degradation due to the data quality of sensors. To overcome the above-mentioned flaws, the future research directions include: ① the multi-agent DRL algorithms will be explored to tackle the extensive and complex power grids; and ② during algorithm training, the data missing situations should be considered. Methods to mitigate the impact of data missing will be adopted to improve the proposed GAT-PPO algorithm, thereby continuously enhancing its robustness. Additionally, measures such as enhancing signal reception strength and improving transmission methods can be adopted to alleviate transmission issues with sensors.

References

[1] B. She, F. Li, H. Cui *et al.*, "Fusion of microgrid control with model-free reinforcement learning: review and vision," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3232-3245, Jul. 2023.

[2] M. Abdelghany, V. Mariani, D. Liuzza *et al.*, "A unified control platform and architecture for the integration of wind-hydrogen systems into the grid," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 4042-4057, Jul. 2023.

[3] Y. Chi, A. Tao, X. Xu *et al.*, "An adaptive many-objective robust optimization model of dynamic reactive power sources for voltage stability enhancement," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 6, pp. 1756-1769, Nov. 2023.

[4] M. Savargaonkar, I. Oyewole, A. Chehade *et al.*, "Uncorrelated sparse autoencoder with long short-term memory for state-of-charge estimations in lithium-ion battery cells," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 15-26, Jan. 2024.

[5] C. Lei, S. Bu, Q. Wang *et al.*, "Look-ahead rolling economic dispatch approach for wind-thermal-bundled power system considering dynamic ramping and flexible load transfer strategy," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 186-202, Jan. 2024.

[6] K. Xie, J. Dong, C. Singh *et al.*, "Optimal capacity and type planning of generating units in a bundled wind-thermal generation system," *Applied Energy*, vol. 164, pp. 200-210, Feb. 2016.

[7] M. Abdelghany, A. Al-Durra, D. Zhou *et al.*, "Optimal multi-layer economical schedule for coordinated multiple mode operation of wind-solar microgrids with hybrid energy storage systems," *Journal of Power Sources*, vol. 591, pp. 1-16, Jan. 2024.

[8] Y. Li, W. Li, W. Yan *et al.*, "Probabilistic optimal power flow considering correlations of wind speeds following different distributions," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1847-1854, Jul. 2014.

[9] Y. Xu, Z. Dong, and R. Zhang, "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4398-4408, Nov. 2017.

[10] M. Lubin, Y. Dvorkin, and S. Backhaus, "A robust approach to chance constrained optimal power flow with renewable generation," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3840-3849, Sept. 2016.

[11] P. Li, C. Zhang, Z. Wu *et al.*, "Distributed adaptive robust voltage-var control with network partition in active distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2245-2256, May 2020.

[12] F. Mráz, "Calculating the exact bounds of optimal values in LP with interval coefficients," *Annals of Operations Research*, vol. 81, pp. 51-62, Jun. 1998.

[13] C. Zhang, H. Chen, Z. Liang *et al.*, "Reactive power optimization under interval uncertainty by the linear approximation method and its modified method," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4587-4600, Sept. 2018.

[14] B. Zhang and Y. Gao, "Data-driven voltage/var optimization control for active distribution network considering PV inverter reliability," *Electric Power Systems Research*, vol. 224, pp. 1-14, Nov. 2023.

[15] K. Xiong, D. Cao, G. Zhang *et al.*, "Coordinated volt/var control for photovoltaic inverters: a soft actor-critic enhanced droop control approach," *International Journal of Electrical Power & Energy Systems*, vol. 149, pp. 1-13, Jul. 2023.

[16] R. Huang, Y. Chen, T. Yin *et al.*, "Learning and fast adaptation for grid emergency control via deep meta reinforcement learning," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4168-4178, Nov. 2022.

[17] Q. Ma and C. Deng, "Simplified deep reinforcement learning based volt-var control of topologically variable power system," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 5, pp. 1396-1404, Sept. 2023.

[18] R. Hossain, Q. Huang, and R. Huang, "Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control," *IEEE Transactions on Power Systems*, vol. 36, no. 5,

pp. 4848-4851, Sept. 2021.

[19] S. Song, Y. Jung, G. Jang *et al.*, "Proximal policy optimization through a deep reinforcement learning framework for remedial action schemes of VSC-HVDC," *International Journal of Electrical Power & Energy Systems*, vol. 150, pp. 1-10, Aug. 2023.

[20] L. Yin, S. Luo, Y. Wang *et al.*, "Coordinated complex-valued encoding dragonfly algorithm and artificial emotional reinforcement learning for coordinated secondary voltage control and automatic voltage regulation in multi-generator power systems," *IEEE Access*, vol. 8, pp. 180520-180533, Oct. 2020.

[21] P. Veličković, G. Cucurull, A. Casanova *et al.*, "Graph attention networks," in *Proceedings of 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, May 2018, pp. 1-12.

[22] E. Vittal, M. O'Malley, and A. Keane, "A steady-state voltage stability analysis of power systems with high penetrations of wind," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 433-442, Feb. 2010.

[23] O. Reyes, C. Morell, and S. Ventura, "Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context," *Neurocomputing*, vol. 161, pp. 168-182, Aug. 2015.

[24] W. Zhang, Z. Wei, B. Wang *et al.*, "Measuring mixing patterns in complex networks by Spearman rank correlation coefficient," *Physica A – Statistical Mechanics and Its Applications*, vol. 451, pp. 440-450, Jun. 2016.

[25] G. Calviño, J. Olivares, and F. Estrada, "Information entropy and fragmentation functions," *Nuclear Physics A*, vol. 1036. pp. 1-17, Aug. 2023.

Xiaofei Liu received the M.S. degree in electrical engineering from Xi'an University of Science and Technology, Xi'an, China, in 2017. He is currently working toward the Ph.D. degree in the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include voltage stability and control of power systems.

Pei Zhang received the Ph. D. degree from Imperial College of Science, Technology and Medicine, University of London, London, UK, in 1999. He is now a Professor at Beijing Jiaotong University, Beijing, China, and he is also a Professor at Tianjin University, Tianjin, China. His research interests include power system operation, power system planning, renewable energy integration, artificial intelligence and its application in power systems.

Hua Xie received the Ph.D. degree from Tsinghua University, Beijing, China, in 2004. She is currently an Associate Professor with Beijing Jiaotong University, Beijing, China. Her research interests include planning and control in active distribution networks.

Xuegang Lu received the M.S. degree from the State Grid Electric Power Research Institute, Nanjing, China. He is currently a Manager of Yunnan Electric Power Dispatching and Control Center, Kunming, China. His research interests include development and construction of power grid dispatching automation system and research and application of digital and intelligent technology in power grid operation and control.

Xiangyu Wu received the Ph.D. degree from Tsinghua University, Beijing, China, in 2017. He is currently an Associate Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include control and stability analysis of microgrids and power-electronic-based power systems.

Zhao Liu received the Ph. D. degree from the State University of New York, Binghamton, USA, in 2020. He is currently an Associate Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include power system transient stability analysis, and applications of machine learning technology in power systems.