

Distribution Network Optimization Based on Topology Security-constrained Integrated Reinforcement Learning

Haixiang Zang, *Senior Member, IEEE*, Yongkai Zhao, Kang Sun, *Student Member, IEEE*, Guoqiang Sun, *Member, IEEE*, Lilin Cheng, *Member, IEEE*, Jingxuan Liu, *Graduate Student Member*, and Zhinong Wei, *Member, IEEE*

Abstract—With the increasing penetration of large-scale renewable energy sources into the power grid, distribution networks are facing significant challenges, including intensified voltage fluctuations and increased network losses. Although deep reinforcement learning has made considerable advancements in addressing optimization problems compared to traditional algorithms, there has been limited focus on enhancing convergence and safety in cooperative optimization scenarios, particularly those involving topological reconstruction. To overcome these challenges, this paper proposes a distribution network optimization model that incorporates topological security-constrained integrated reinforcement learning. The model improves the encoding of topologies by representing them in a multi-dimensional discrete space and introduces a topological masking mechanism to achieve high safety and computational efficiency. Additionally, an ensemble strategy is utilized to develop an action network group, improving action prediction and screening, thereby achieving better training stability. Experiments conducted on an enhanced IEEE33-node distribution network system indicate that the proposed improvements significantly enhance training stability and support the safe and efficient operation of the system.

Index Terms—Distribution network optimization, ensemble learning, reinforcement learning, topology reconstruction.

I. INTRODUCTION

With the gradual reduction of fossil fuel reserves, the push for a low-carbon energy transition has

attracted widespread attention from the international community, promoting the extensive integration of large-scale renewable and distributed energy sources into power grids to meet growing local energy demands [1]–[3]. Nevertheless, the inherent randomness and volatility of renewable energy generation often pose great challenges, such as system voltage deviations and changes in power flow, presenting substantial obstacles to the secure and stable operation of distribution network systems [4].

To effectively overcome the challenges brought by the uncertainty of renewable energy sources, existing studies on dispatch optimization emphasize the integration of various schedulable resources on both generation and load sides, including energy storage systems, reactive power compensation devices, as well as flexible loads, while incorporating dynamic reconfiguration strategies on grid side [5]–[7]. This integrated method aims to achieve coordinated system optimization by directly controlling power flow distribution within the network. Traditional research frameworks typically rely on mathematical modeling and optimization algorithms, especially robust optimization and dynamic programming methods that employ convex relaxation or linearization of power system models to provide stable and reliable solutions for general scenarios [8]. Nevertheless, when dealing with complex optimization problems under uncertain conditions, the computational efficiency of these traditional optimization algorithms is significantly reduced, making it difficult to meet the strict requirements of real-time power grid scheduling. Although heuristic algorithms, such as particle swarm optimization and genetic algorithms, demonstrate certain search capabilities in high-dimensional spaces, their solution time still increases dramatically with the complexity of the problem, and they have limited ability to ensure global optimality.

In this context, deep reinforcement learning (DRL) theory has become a promising approach for optimizing the dispatch of complex power grids. A plethora of studies have investigated the development of optimization

Received: May 4, 2025

Accepted: December 5, 2025

Published Online: March 1, 2026

Haixiang Zang (corresponding author), Yongkai Zhao, Kang Sun, Guoqiang Sun, Lilin Cheng, Jingxuan Liu, and Zhinong Wei are with the School of Electrical and Power Engineering, Hohai University, Nanjing 210098, China (e-mail: zanghaixiang@hhu.edu.cn; hhu_zyk@hhu.edu.cn; sunkang.real@outlook.com; hhusunguoqiang@163.com; straw@hhu.edu.cn; hhluljx98@hhu.edu.cn; wzn_nj@263.net).

DOI: 10.23919/PCMP.2024.000440

frameworks based on DRL [9]. Reference [10] uses deterministic policy gradient algorithms (DDPG) with a bi-level optimization structure to address economic dispatch and reconfiguration challenges, while reference [11] demonstrates strong control over voltage violations by incorporating voltage sensitivity analysis into gradient calculations. Reference [12] uses DRL to simulate the mixed integer linear programming (MILP) solver instead of from-scratch training, which can accurately determine the optimal load energy allocation strategy. These methods do not rely on precise system parameters or physical models. On the contrary, they can quickly adapt to environmental changes by extracting and representing the key characteristics and knowledge models of power systems. This capability enables them to train general decision-making models to directly solve complex optimization problems, avoiding the need for single-use strategies and achieving extremely fast solution times [13].

Although reinforcement learning (RL) has impressive decision-making efficiency in managing the complexity of dimensions, few studies have directly applied it to optimization scenarios such as distribution network reconfiguration (DNR) [14] or distribution network fault restoration [15]. In the domain of topology optimization for complex power systems, effectively characterizing the topological action space is a key challenge in RL applications. System topology reconstruction is essentially a constrained combinatorial optimization problem, where the solution space includes a large number of infeasible topologies that violate physical constraints (such as loop structures and islanded operation states in power systems). To convert the discrete combinatorial optimization problem to a form amenable to RL, topology encoding mechanisms map system topology information into a finite-dimensional parameter space [16]. The design of these encoding mechanisms directly affects the optimization dimensionality and computational efficiency of the model. Binary encoding based on traditional optimization algorithms meets the requirements of dimensionality reduction. Nevertheless, its high-dimensional sparsity results in effective actions occupying only the smallest portion of the search space, which reduces the validity of topological decisions. One-hot encoding, derived from filtering valid topology sets, ensures action validity, but it faces challenges such as dimensionality explosion and the disruption of topological adjacency relationships. The continuous space-based encoding mechanism achieves topological decision-making through interval discretization but is highly sensitive to parameter perturbations [17]. Reference [18] proposes using a topology mask to map the initial topology, but the computation of the mask elements relies on general connectivity detection algorithms, and exhaustive enumeration of the action space

requires massive computations, which to some extent limits the generalization of the method to practical application scenarios. These contradictions highlight the necessity of topology encoding mechanisms to address the following key elements: 1) limiting encoding dimensionality to avoid the curse of dimensionality; 2) ensuring encoding space continuity, where similar encodings correspond to adjacent topologies in the electrical characteristic space; and 3) establishing rigorous mathematical mappings to ensure the integrity and validity of the action space while maintaining computational efficiency. Although significant progress in topology encoding design, existing studies still have shortcomings in systematically integrating these elements to construct a comprehensive encoding mechanism.

Additionally, when dealing with large-scale optimization problems such as distribution network reconfiguration, the learning process may be very sensitive to the action trajectories guided by the initial parameters. This sensitivity may not only reduce learning efficiency but also cause the formation of inherent biases in the policy network at the early stage of training, thereby increasing the risk of the model falling into local optima [19]. To address this issue, researchers have attempted to improve the performance of DRL algorithms by optimizing the decision structure [20]. Reference [21] applies noisynet to deep Q-network (DQN), automatically adjusting the exploration efficiency of the topology during training by adding perturbations to the network weights to alleviate the adverse effects of large-scale action spaces on intelligent learning performance. Reference [22] introduces the concept of “switch contribution” and constructs a multilayer learning architecture based on a reward-sharing mechanism, which removes low-contributing branch switches in the preliminary reconstruction process by quantizing and assigning multi-intelligence weights, gradually accelerating the intelligence training process. Following the principle of imitation learning, a safe RL method is proposed based on tutorial learning to help accelerate and fit the cases in the expert knowledge base to construct a tutorial learning model [23]. By providing scheduling decision guidance to the intelligences, it significantly improves the convergence efficiency of RL and the safety of the initial decision-making. Nevertheless, the multi-intelligence training approach requires more computational resources, and the cost of collecting grid expert strategy data is also extremely high. As a result, under the constraints of limited computational and data resources, it is crucial to consider constructing algorithms that can effectively guide the decision-making of distribution grid reconfiguration at the early stage of training.

To overcome these challenges, this paper proposes a power system dispatching strategy that integrates

topology safety constraints with RL. The contributions of this paper are as follows.

1) A low-dimensional safety-aware topology encoding method is proposed to reduce the significant risks brought by out-of-limit actions generated by RL models in power systems. This method introduces a masking mechanism to dynamically prune branches in each cycle to generate tree-like topologies. Meanwhile, reparameterization techniques are utilized to improve the practicality of the encoding.

2) A specialized topology detection algorithm and hash mapping strategy are put forward to meet the strict computational efficiency requirements of real-time optimization in distribution network reconfiguration. Compared with traditional topology connectivity detection algorithms, this method shows significantly higher computational efficiency.

3) An integrated mechanism for policy optimization in complex scenarios is developed. To overcome the convergence challenges of policy optimization in complex distribution network scenarios, an integrated mechanism integrating multi-network prediction, soft parameter updates, and network pruning is devised. This mechanism ensures the rationality of initial action policies while occupying minimal computational resources, leading to higher training efficiency of the optimization model.

The structure of the remaining part of this paper is organized as follows. Section II introduces the computational principles of topological coding and security constraints, as well as an improved reinforcement learning algorithm. Section III constructs a Markov decision-making environment for collaborative optimization of distribution networks and presents the corresponding constraints. Section IV demonstrates the experimental parameters and application framework, evaluating the proposed method. Section V provides the main conclusions of this paper.

II. TOPOLOGICAL SECURITY-CONSTRAINED INTEGRATED RL MODEL

A. Topological Encoding and Security Constraints

In traditional DNR, a “closed-loop design and open-loop operation” approach is usually employed to balance economy and security. The dimension explosion problem due to the increase in the number of its system branches and the radial topological constraints in the operation process make it highly challenging for the existing RL techniques to solve optimization problems involving discrete topological spaces [24]. In contrast, integrating domain knowledge and priori information to encode the distribution network topology can structure the solution space of switching states and enhance the solution efficiency of optimization algorithms. For this reason, this paper aims to construct a novel topology coding mechanism for distribution networks based on loop coding [25]. Loop coding, which is investigated in terms of basic loops, inherently correlates topological similarity with coding similarity

and can greatly reduce the topological space dimension. Based on this, topology masks are introduced to map multiple discrete topologies into a secure multidimensional discrete space, but the computation of the mask inevitably increases the computational burden of the optimization solution. Considering the computational efficiency of topological coding, this paper further devises a unique connectivity detection algorithm following the branch-switching principle, which substantially improves the computational efficiency of the mask elements. Additionally, a hash mapping strategy is embedded to achieve high timeliness of the solution under sufficient storage space.

According to graph theory, when all contact switches in a distribution network are closed, the topology of the network consists of a series of interconnected independent loops. A necessary but not sufficient condition for the transition from a ring topology to a radiant topology is that at least one branch switch must be activated in each independent loop. Once this condition is satisfied, the model can be configured to selectively disconnect a controllable branch in each different independent loop in sequence. Let N represent the total number of branches; N_{switch} represent the total number of controllable switches; N_L represent the total number of independent loops; the number of controllable switches within each independent loop is denoted as $N_{\text{switch},l}$; the total number of possible solutions is N_{topo} ; and the corresponding multidimensional discrete action space is $\mathcal{A}_{\text{topo}}$, which can be given by:

$$\mathcal{A}_{\text{topo}} = \{\mathbf{a}_l | l \in \mathcal{L}\} \quad (1)$$

$$\mathbf{a}_l = [a_{l,1}, a_{l,2}, \dots, a_{l,i}]_{i \in \mathcal{E}_{\text{switch},l}, a_{l,i} \in \{0,1\}} \quad (2)$$

$$N_{\text{topo}} = \prod N_{\text{switch},l} \quad (3)$$

where \mathcal{L} denotes the set of independent loop indices; $\mathcal{E}_{\text{switch},l}$ represents the set of all controllable switch indices within the l th independent loop; vector \mathbf{a}_l stands for the discrete action executed within the l th independent loop; and $a_{l,i}$ denotes the action code for the controllable switch in the l th independent loop, with 0 indicating that the switch is connected and 1 indicating that the switch is disconnected.

This method inherently adheres to certain topological constraints, thereby improving the efficiency of obtaining feasible solutions. Nevertheless, it may not always meet the requirements for radial topology in arbitrary scenarios. To ensure operational safety and effectively reduce the operation space, this paper introduces the concept of action masks to eliminate illegal operations that violate topological constraints. The computational framework is illustrated in Fig. 1, and the following equations are provided:

$$\mathbf{m}_l = [m_{l,1}, m_{l,2}, \dots, m_{l,N_{\text{switch},l}}]_{m_{l,i} \in \{0, -\infty\}} \quad (4)$$

$$\mathbf{a}_l^{\text{out}} = [a_{l,1}^{\text{out}}, a_{l,2}^{\text{out}}, \dots, a_{l,N_{\text{switch},l}}^{\text{out}}]_{a_{l,i}^{\text{out}} \in \mathcal{R}} \quad (5)$$

$$\mathbf{a}_l^{\text{mask}} = \mathbf{a}_l^{\text{out}} + \mathbf{m}_l = [a_{l,1}^{\text{mask}}, a_{l,2}^{\text{mask}}, \dots, a_{l,N_{\text{switch},l}}^{\text{mask}}] \quad (6)$$

where \mathbf{m}_l represents the mask vector for loop l , with $m_{l,i}$ denoting the mask for branch switch i within this loop; $a_{l,i}^{\text{out}}$ and $a_{l,i}^{\text{mask}}$ stand for the original data and masked action data for branch switch i in this loop, respectively; $\mathbf{a}_l^{\text{out}}$ denotes the original output from the action network; and $\mathbf{a}_l^{\text{mask}}$ denotes the masked action output.

$$g_i = -\log(-\log(u_i)), u_i \sim U(0,1) \quad (7)$$

$$a_{l,i} = \frac{\exp\left(\frac{g_i + \log a_{l,i}^{\text{mask}}}{\tau}\right)}{\sum_{j=1}^{N_{\text{switch},l}} \exp\left(\frac{g_j + \log a_{l,j}^{\text{mask}}}{\tau}\right)} \quad (8)$$

where reparameterization techniques are utilized to randomly sample the action outputs to ensure consistency in action probabilities, with g_i representing the random distribution noise [26]; u_i is a random variable following a uniform distribution; and τ represents the temperature coefficient.

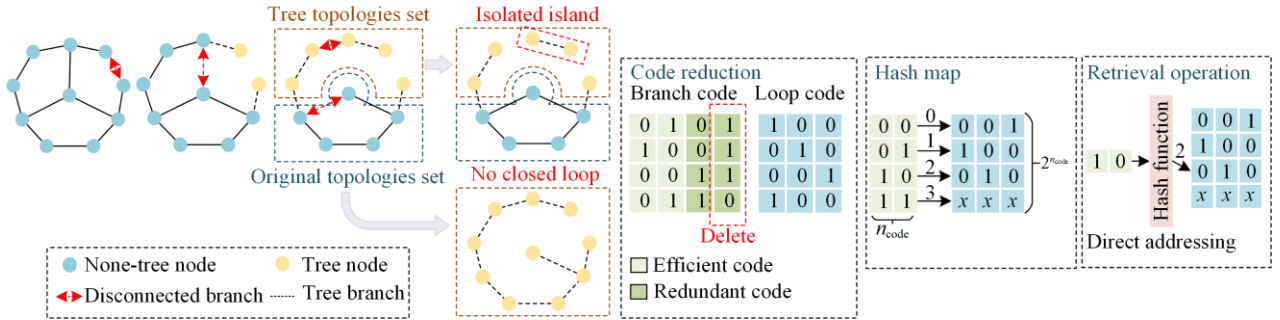


Fig. 2. The structure of the topology detection algorithm.

The new detection algorithm regards topology generation as the process of extracting the tree topology from the original closed topology. Its core feature is that once tree nodes and tree branches are constructed, they become part of the final tree topology. Disconnecting any tree branch will inevitably form an island, eliminating the need for repeated connectivity checks for tree branches. Moreover, with islands, a closed-loop topology cannot be formed, ensuring that the tree topology has no islands or closed loops. Therefore, this mechanism simplifies the tree topology constraint of the grid into a constraint on tree branch connectivity. The action mask only needs to ensure that tree branches are not disconnected by subsequent loops. The algorithm, by dynamically updating all branch categories, can determine all mask elements for the corresponding loop in a single pass, thereby avoiding redundancy.

Without loss of generality, the general pseudocode of the algorithm is presented in Algorithm 1.

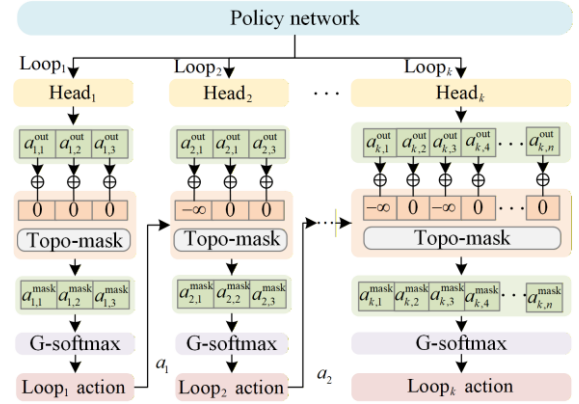


Fig. 1. Topology mask generation framework.

Masked elements are acquired through connectivity detection by traversing branches, which has a large impact on model training efficiency. To address this issue, this paper proposes a more efficient detection algorithm to rapidly identify topological constraints.

As shown in Fig. 2, a tree node is defined as a node where the number of connected non-tree branches is less than or equal to one. A tree branch is characterized as a branch containing a tree node. Initially, all branches are characterized as non-tree branches.

Algorithm 1 Topological Mask Action Generation

1. Initialize temporary node set $\mathcal{T} \leftarrow \emptyset$, tree branch set $\mathcal{B}_{\text{tree}} \leftarrow \emptyset$
 2. $a_0 = \text{Sample}(\varepsilon_{\text{switch},0})$; sample controllable branch elements a_0 from the loop set $\varepsilon_{\text{switch},0}$ and store them as (i_0, j_0)
 3. **for** $n=0$ to $|\mathcal{L}|$ **do**
 4. Add branch elements a_n to the end of list A; $\mathcal{B} \leftarrow \mathcal{B} - \{a_n\}$; $\mathcal{T} \leftarrow \mathcal{T} \cup \{i_n, j_n\}$
 5. **while** $|\mathcal{T}| \geq 1$ **do**
 6. Take any node element k from node set \mathcal{T} , $\mathcal{T} \leftarrow \mathcal{T} - \{k\}$, calculate the total number N_{connect} of branches in set \mathcal{B} containing node k
 7. **if** $N_{\text{connect}} = 1$ **then**
 8. Use (v, k) to represent this branch; $\mathcal{B} \leftarrow \mathcal{B} - \{(v, k)\}$; $\mathcal{B}_{\text{tree}} \leftarrow \mathcal{B}_{\text{tree}} \cup \{(v, k)\}$; $\mathcal{T} \leftarrow \mathcal{T} \cup \{v\}$
 9. **end if**
 10. **end while**
 11. Mask the branch elements contained in list A or set \mathcal{B} in $\varepsilon_{\text{switch},n+1}$, and sample the branch element a_{n+1} to be disconnected from the next loop
 12. **end for**
 13. **return** A
-

Since loop encoding inherently compresses the topological space effectively and some loop encodings exhibit decoupling relationships, a hash mapping strategy can be naturally established to store the topology mask, thereby further improving the computational efficiency of the algorithm.

$$f^{\text{mask},l} : \mathbf{B}_{n_{\text{switch}} \times N_{\text{switch}}^l} \rightarrow \mathbf{M}_{n_{\text{switch}} \times N_{\text{switch}}^l} \quad (9)$$

where N_{switch}^l denotes the total number of switches that can operate before the current loop; and n_{switch} represents the total number of valid topologies that can be generated by these switches; matrix \mathbf{B} stores all valid states in multi-hot encoding; while matrix \mathbf{M} stores the corresponding mask elements; function $f^{\text{mask},l}$ defines a mapping process that transforms the corresponding operation encoding into a set of mask vectors, it traverses the action space using the aforementioned algorithm and randomly stores the results into the matrices \mathbf{B} and \mathbf{M} .

The data volume of this type of hash structure increases with the number of controllable branches, accompanied by a certain degree of spatial redundancy. To alleviate this issue, this paper preemptively prunes the state encoding and the calculations are given by:

$$R(\mathcal{E}_{\text{select},l}) = \forall i, j \in \{1, 2, \dots, n_{\text{switch}}\} (i \neq j), \quad (10)$$

$$\text{s.t. } \mathbf{B}_{\text{select},l,i} = \mathbf{B}_{\text{select},l,j} \text{ and } \mathbf{M}_{l,i} = \mathbf{M}_{l,j}$$

$$\mathbf{T}_l = [b_{i,k}]_{n_{\text{switch}} \times (N_{\text{switch}}^l - |\mathcal{E}_{\text{select},l}|)} \quad (11)$$

where $R(\cdot)$ denotes a necessary condition for the existence of the mapping; $\mathcal{E}_{\text{select},l}$ is a set, and each column of matrix \mathbf{B} is encoded into this set, resulting in matrix $\mathbf{B}_{\text{select},l}$; $\mathbf{M}_{l,i}$ and $\mathbf{M}_{l,j}$ are the original corresponding mapping values; \mathbf{T}_l stores the mask vectors for all valid states; and $b_{i,k}$ is the row vector element of matrix \mathbf{T}_l . If the necessary condition $R(\cdot)$ is met, the branch is considered redundant and its encoding is added to the set $\mathcal{E}_{\text{select},l}$; otherwise, the encoding is eliminated from the set, the state data of the removed branch are restored, and the next branch encoding is chosen for inclusion. This process is repeated until all columns are traversed, leading to the formation of the reduced mapping matrix \mathbf{T}_l .

$$h_{\text{mask}}(b_i) = \sum 2^{k-1} b_{i,k} \quad (12)$$

$$\text{mask}(b_i) = \mathbf{M}_{h_{\text{mask}}(b_i)} \quad (13)$$

In this process, b_i is inserted into the hash function $h_{\text{mask}}(\cdot)$ to obtain the index remapping, thereby rapidly calculating the mask vector for the current independent loop; in this context, the hash function is implemented using direct addressing, achieving a balance between

space complexity and time complexity for the corresponding hash table \mathbf{H}_l .

B. Twin Delayed Deep Deterministic Policy Gradient Algorithm Based on Ensemble Learning

The twin delayed deep deterministic policy gradient (TD3) algorithm is designed based on the actor-critic (AC) architecture. By incorporating dual networks, policy smoothing, and delayed updating mechanisms, it has significantly improved the handling of low-dimensional continuous control problems across various power system applications [27]. However, its suboptimal performance in high-dimensional topological spaces makes it difficult to achieve effective topological reconfiguration in distribution networks. This limitation is due to the constraint of randomly initialized parameters on the agent. Even if an action is feasible at a certain moment, the agent may still deviate from suboptimal action trajectories over time, leading to extensive ineffective exploration in the early stages and delaying convergence to the optimal strategy [28].

To deal with this issue, this paper introduces an ensemble strategy to predict outcomes before executing actions, thereby improving overall exploration efficiency through prior action selection. As demonstrated in Fig. 3, action networks with identical structures but different initialization parameters are stacked within the policy network. Taking the current state as input, varying scheduling policies are generated and transmitted to a parallel pseudo-environment group. Then, the optimal action is selected and applied to the actual power grid environment. Although the pseudo-environment is designed to mimic the real environment, it can be simplified to improve computational efficiency, as the algorithm aims to discard poor scheduling policies instead of calculating precise reward values. In this paper, distributed power flow is utilized in place of AC power flow to rapidly compute the steady state of the system.

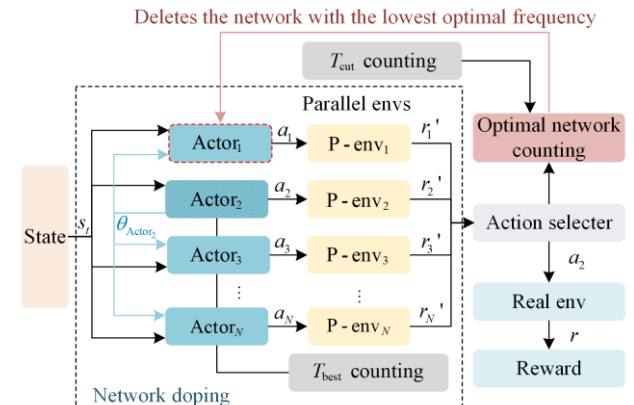


Fig. 3. The structure of the ensemble learning (EL)-TD3 network.

Due to random parameter initialization, certain actor-networks may always perform suboptimally in specific scenarios, significantly decreasing the efficiency of action selection. To fully utilizing decision data while perserving both the uniqueness and diversity of network parameters, a parameter perturbation mechanism is introduced to adjust the parameters of suboptimal models. The algorithm utilizes a period counter to track the number of optimal actions of each actor-network over a defined period T_{best} . At the end of this cycle, the parameters of the best-performing actor-network undergo soft updates across the other actor-networks, guiding them to achieve optimal performance while maintaining diversity. The equation governing this process is given by:

$$\phi_{\text{others}} = \tau_{\text{best}} \phi_{\text{best}} + (1 - \tau_{\text{best}}) \phi_{\text{others}} \quad (14)$$

where ϕ_{others} denotes the parameters of the actor-networks excluding the optimal network; ϕ_{best} represents the parameters of the optimal actor-network; and τ_{best} indicates the soft update coefficient.

For each actor-network, parameter updates rely only on the value network's output for gradient ascent optimization. Considering that all actor-networks have the same value network, without appropriate intervention, the actions produced by the policy network tend to converge as the training process continues. Since the likelihood of suboptimal scheduling strategies decreases over time and the soft update method affects the convergence of model parameter training, a policy network reduction mechanism is further introduced to guarantee that the ensemble policy mainly operates during the early stages of training. This mechanism tracks the frequency at which each network is selected as the optimal network over an extended period T_{cut} . At the end of each period, the actor-network with the lowest selection frequency is eliminated. This process continues until only one actor-network remains, returning the architecture to the original TD3 framework. This method effectively minimizes redundancy, refines the policy network structure, and significantly enhances computational performance.

The pseudocode of the complete training process under the action integration mechanism is presented in Algorithm 2.

Algorithm 2 EL-TD3

1. Initialize critic networks Q_{θ_1} , Q_{θ_2} , and actor networks $\pi_{\phi_{l_n}}$ with random parameters θ_1 , θ_2 , ϕ_{l_n} , initialize target networks $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$, $\phi' \leftarrow \phi$
 2. Initialize list ε as a sequence from 1 to n , optimal identification bit $i_{\text{best}} \leftarrow 1$, optimal counting parameter $N_{\text{best}} \leftarrow 0$, $N_{\text{cut}} \leftarrow 0$, reset the relay buffer B , environment E , and pseudo-environment group E_{l_n}'
 3. **for** $n=1$ to T **do**
 4. Select action with exploration noise $a_{l_n} \sim \pi_{\phi_{l_n}}(s) + \varepsilon$, $\varepsilon \sim N(0, \sigma)$, and based on the reward r_{l_n}' from the pseudo-environment group, select the action a_{best} corresponding to the highest reward, observe the reward r and new state s' from the real environment, store them in the experience buffer B in the form (s, a, r, s') , and update the index i_{best} using the optimal network, $N_{\text{best}, i_{\text{best}}} \leftarrow N_{\text{best}, i_{\text{best}}} + 1$
 5. Sample mini-batch of N transitions (s, a, r, s') from B
 6. $\tilde{a} \leftarrow \pi_{\phi'}(s') + \varepsilon$, $\varepsilon \sim \text{clip}(N(0, \bar{\sigma}), -c, c)$
 7. $y \leftarrow r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a})$
 8. **if** $t \bmod d$ **then**
 9. **for** i in ε **do**
 10. $\nabla_{\phi} J = N^{-1} \sum_a \nabla_a Q_{\theta'_i}(s, a) \Big|_{a=\pi_{\phi}(s)}$; update ϕ by network policy gradient
 11. **end for**
 12. $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$, $\phi' \leftarrow \tau\phi_{\text{best}} + (1-\tau)\phi'$
 13. **end if**
 14. **if** $t \bmod T_{\text{best}}$ **then**
 15. Select optimal params ϕ_{best} based on N_{best}
 16. $\phi_{\text{others}} \leftarrow \tau_{\text{best}} \phi_{\text{best}} + (1 - \tau_{\text{best}}) \phi_{\text{others}}$, $N_{\text{cut}, i_{\text{best}}} \leftarrow N_{\text{cut}, i_{\text{best}}} + 1$
 17. **end if**
 18. **if** $t \bmod T_{\text{cut}}$ **then**
 19. Select the worst action network based on N_{cut} , and remove it from the list ε
 20. **end if**
 21. **end for**
-

III. MODEL FOR DYNAMIC DISTRIBUTION NETWORK RECONFIGURATION CONSIDERING PHOTOVOLTAICS AND ENERGY STORAGE

In distribution systems with high penetration rates, it is a common and practical solution to use topology reconstruction technology and energy storage devices for energy management. However, research often overlooks the uncertainty of scenarios. To achieve dynamic coordination between topology and energy storage, this paper considers the fluctuation characteristics of photovoltaics and loads, thereby providing reliable and effective scheduling solutions for operators.

A. Markov Decision Model

The optimization problem of dynamic distribution network reconfiguration is formulated as a Markov decision environment that can be solved using RL. To

model this environment, various components are described in detail as follows.

1) Action

In this optimization problem, the model actions include system topology reconstruction and adjustment of energy storage charging and discharging strategy.

$$\mathbf{A}_t = \{\mathbf{P}_t^{\text{ess}}, \mathbf{A}_{\text{topo}}\} \quad (15)$$

where $\mathbf{P}_t^{\text{ess}}$ denotes the charging and discharging power of the energy storage device at time t , with a positive value indicating discharging and a negative value indicating charging; \mathbf{A}_{topo} represents the topological action of the model at time t .

2) State

The state space defined in this paper includes:

$$\mathbf{S}_t = \{\mathbf{P}_t^{\text{pv}}, \mathbf{G}_t, \mathbf{E}_t\} \quad (16)$$

where \mathbf{S}_t represents the system state space at time t ; \mathbf{P}_t^{pv} denotes the vector of PV power output from the previous period; \mathbf{G}_t indicates the system topology code; and \mathbf{E}_t reflects the current energy storage capacity.

3) Reward

The reward function evaluates both the upper-level power purchase cost of the system during the day-ahead scheduling period and the cost associated with switching times in the network reconstruction process. The reward is represented by:

$$\begin{cases} r_t = -\sum_{i=1}^L (c_t^{\text{pur}} P_t^{\text{pur}} + c_{\text{sw}} n_{\text{opr}}) + k n_{\text{const}} + b \\ P_t^{\text{pur}} = P_t^{\text{loss}} + P_t^{\text{load}} + P_t^{\text{ess}} - P_t^{\text{pv}} \\ P_t^{\text{ess}} = \sum_{j=1}^{N_{\text{ess}}} \left(\frac{1/\eta_{c,j} - \eta_{d,j}}{2} P_{j,t}^{\text{ess}} - \frac{1/\eta_{c,j} + \eta_{d,j}}{2} |P_{j,t}^{\text{ess}}| \right) \end{cases} \quad (17)$$

where L represents the length of the scheduling period, which is 24 h; c_t^{pur} is the power purchase cost parameter from the superior grid; while P_t^{load} , P_t^{pv} , and P_t^{loss} respectively indicate the load power, PV output, and system network loss at time t ; N_{ess} denotes the number of energy storage devices; while $\eta_{c,j}$ and $\eta_{d,j}$ are the charging and discharging coefficients of the j th energy storage device, respectively; c_{sw} is the cost coefficient associated with switching actions; and n_{opr} represents the total number of switching operations at time t ; additionally, n_{const} represents the number of violations of constraints, with k being the corresponding penalty coefficient; and b is a bias parameter for reward shaping; $P_{j,t}^{\text{ess}}$ represents the output of the j th energy storage at time t .

B. Constraints

The constraint factors include topological constraints, energy storage constraints, and power flow constraints, with topological radial constraints being guaranteed by the security layer. Nevertheless, due to the wear and tear of branch switches during topology reconstruction, it is necessary to limit the number of switch operations within a reasonable range.

$$\begin{cases} N^{\text{opr}} = \sum_{t=1}^L \sum_{m=0}^M |S_{m,t}^{\text{sw}} - S_{m,t-1}^{\text{sw}}| \leq N_{\text{max}}^{\text{opr}} \\ N_m^{\text{opr}} = \sum_{t=1}^L |S_{m,t}^{\text{sw}} - S_{m,t-1}^{\text{sw}}| \leq N_{\text{max},m}^{\text{opr}} \end{cases} \quad (18)$$

where N^{opr} represents the total number of actions performed by all switches during the scheduling period; N_m^{opr} denotes the number of actions for switch m within that period; $S_{m,t}^{\text{sw}}$ represents the state of switch m at time t , where 0 indicates that the switch is on and 1 indicates that the switch is off; $N_{\text{max}}^{\text{opr}}$ denotes the maximum allowable number of actions for all switches during the scheduling period; and $N_{\text{max},m}^{\text{opr}}$ denotes the maximum allowable number of actions for switch m .

Energy storage constraints and power flow constraints are defined as follows [29]:

$$\begin{cases} P_j^{\text{ess,min}} \leq P_{j,t}^{\text{ess}} \leq P_j^{\text{ess,max}} \\ E_{t+1} = E_t - \frac{1/\eta_d + \eta_c}{2} P_t^{\text{ess}} - \frac{1/\eta_d - \eta_c}{2} |P_t^{\text{ess}}| \\ E_j^{\text{min}} \leq E_{j,t} \leq E_j^{\text{max}} \\ 0.1 \leq \delta_{\text{SOC},t} \leq 0.9 \end{cases} \quad (19)$$

where $P_j^{\text{ess,min}}$ and $P_j^{\text{ess,max}}$ are the minimum and maximum limits of energy storage charging and discharging power, respectively; E denotes the energy storage capacity; E_j^{min} and E_j^{max} are the lower and upper limits for the j th energy storage, respectively; and $\delta_{\text{SOC},t}$ represents the state of charge (SOC) of the stored energy at time t .

$$\begin{cases} P_{i,t} = V_{i,t} \sum_{j \in \Omega} V_{j,t} (G_{i,j} \cos \theta_{i,j} + B_{i,j} \sin \theta_{i,j}) \\ Q_{i,t} = V_{i,t} \sum_{j \in \Omega} V_{j,t} (G_{i,j} \sin \theta_{i,j} - B_{i,j} \cos \theta_{i,j}) \\ V_{\text{min}} \leq V_{i,t} \leq V_{\text{max}} \end{cases} \quad (20)$$

where $P_{i,t}$ and $Q_{i,t}$ denote the active and reactive power injected at nodes, respectively; $G_{i,j}$ and $B_{i,j}$ represent the branch conductance and susceptance between nodes i and j , respectively; $\theta_{i,j}$ is the corresponding phase angle between nodes i and j ; $V_{i,t}$ is the voltage at node i at time t , with V_{max} and V_{min} representing the upper and lower limits of the voltage, respectively.

IV. CASE ANALYSIS

A. Experimental Setup

In this study, the enhanced IEEE33-node system is taken as a reference case for simulation. The system incorporates distributed PV generation at 16 nodes to model the uncertainty factors associated with new energy sources. Meanwhile, it consists of 3 energy storage devices and 12 branch switches to realize secure and economical system operation. The PV power generation data are based on typical summer conditions, and additional proportional noise is applied to satisfy the scale requirements for RL training. The noise characteristics are modeled based on measured PV data from a region in Belgium, to make the simulated environment accurately reflect real-world conditions [30]. The detailed parameters for the power system equipment and models are listed in Table I.

TABLE I
MODEL PARAMETER

Parameter	Value	Parameter	Value
PV distribution α	0.778	Discount rate γ	0.99
PV distribution β	0.854	Soft update τ_{critic}	0.05
Hidden layer size	128	Soft update τ_{best}	0.1
Batch size	64	T_{best}	120
Learning rate l_{critic}	0.001	T_{cut}	1200
Learning rate l_{actor}	0.0001	N_{ator}	5
ES capacity (MW)	0.8/0.6/0.4	Buffer size	20 000
ES charging efficiency	0.95	ES charging efficiency	0.95

For the hyperparameters in the reward function, c_{sw} comprehensively accounts for the actual switching losses and operational costs; it is normalized to the same magnitude as the electricity purchase cost, and its value is set to 0.004. The constraint penalty coefficient k and the bias coefficient b are determined using a grid search experiment, with the former being set to -0.5 and the latter being set to 2. A parameter sensitivity heatmap was utilized to visualize the impact of the joint variation of these hyperparameters on the optimization cost, as illustrated in the following Fig. 4.

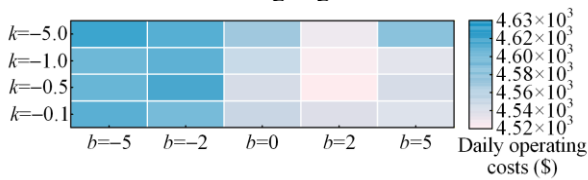


Fig. 4. Hyperparameter grid search heatmap.

As shown in Fig. 4, the optimization performance increases significantly with the bias coefficient. This phenomenon is due to the inherent cost associated with user load in the power grid operation reward function. An appropriate bias guides the reward function to focus more on the relative differences between states and actions, thus balancing the number of positive and negative samples and ensuring precise convergence of model parameters during the training process. And in

addition, under the same bias coefficient, as the penalty coefficient decreases, the agent model is better at capturing constraint correlations in system operations, thereby improving the effectiveness of strategy optimization. Nevertheless, when the penalty coefficient drops below a certain threshold, the optimization cost begins to increase. This is because a too small penalty coefficient leads the model to focus too much on the feasibility of decisions while ignoring other critical optimization factors.

Figure 5 visually demonstrates the embedding mechanism of the proposed algorithm in the actual model training process. When there are sufficient storage resources are available, a hash mapping strategy is employed for rapid lookup, while in lightweight scenarios, the proposed topology mask dynamic generation algorithm is used for computation. Network pruning and parameter update operations are triggered and executed by a periodic timer. The overall framework of the model network is depicted in Fig. 6. The hardware platform is equipped with an Intel Core i7-9750H processor and an NVIDIA RTX 2060 6GB graphics card. The RL model is developed using Python 3.6 with the Tianshou 1.0 framework. In the experiments, AC power flow calculations are conducted to simulate the actual operating environment.

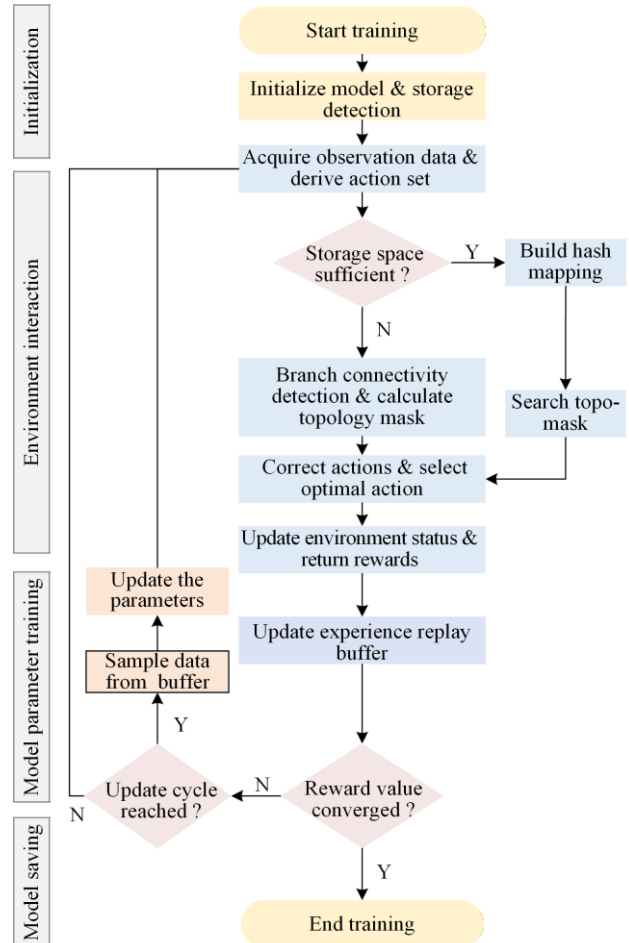


Fig. 5. Algorithm implementation flowchart.

C. Performance Analysis of the Improved EL-TD3 Algorithm

To validate the effectiveness of the improved EL-TD3 algorithm, it is compared with TD3 and DDPG algorithms in solving the cooperative scheduling problem of PV systems [31]. All models are trained five times using fixed random seeds, and the average rewards for every 1000 interaction rounds are recorded. The results are illustrated in Fig. 8.

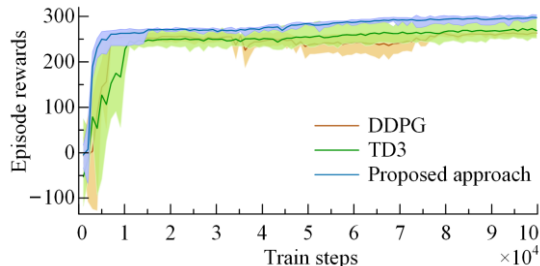


Fig. 8. Comparison of reward for different algorithms.

Figure 8 compares the performance of the proposed algorithm in learning scheduling policies with other algorithms in the same scenario. In Fig. 8, the solid line represents the average rewards, while the shaded area indicates the range of reward fluctuations. It can be seen that the proposed algorithm demonstrates excellent training efficiency, especially in the early stages, owing to the ensemble strategy for a prior action screening. This strategy ensures consistent selections of actions from a better-performing range, accelerating the learning process. Additionally, the proposed model tends to converge after approximately 10 000 interactive training sessions, earlier than the other models. Although the improved algorithm initially requires more computing power, its faster training speed compensates for this. By contrast, the TD3 and DDPG algorithms sometimes exhibit comparable performance to the proposed algorithm, but their reward values fluctuate greatly, and the final reward is usually lower. The emergence of this discrepancy is due to the fact that these models rely on random strategies in certain experiments, which may temporarily remain consistent with the environment, producing favorable results. Nevertheless, as the environment updates and state shifts, the quality of subsequent

actions deteriorates, resulting in suboptimal trajectories. These poor trajectories will be trained repeatedly, making it difficult for the model to find the optimal path, often leading to local optima.

Table III highlights the statistical indicators of the dispersion of reward values for the three algorithms, showing the reliability and convergence characteristics of the proposed algorithm during the training process. The standard deviation of the proposed algorithm is 6.05, substantially lower than that of the other algorithms, demonstrating better consistency. The kurtosis is 1.07, indicating that the reward values are a more concentrated distribution around the mean, with a steeper peak compared with a normal distribution. It is interesting that although the TD3 algorithm shows a higher standard deviation than DDPG, this is not an indicator of poorer performance. On the contrary, it reflects TD3's more precise fitting and stronger capability to explore random noise, enabling the model to discover optimal actions through gradual exploration in certain experiments. By contrast, DDPG always remains within the suboptimal action space, resulting in more reliable but lower overall rewards. The proposed algorithm, however, achieves both reliable convergence and optimal results, outperforming the other algorithms without such trade-offs.

TABLE III
REWARD FLUCTUATION ANALYSIS

Algorithm	Reward		
	Standard deviation	Skewness coefficient	Peak value
DDPG	13.91	-0.19	-1.22
TD3	21.74	0.50	-0.92
MP-TD3	6.05	-0.64	1.07

D. Model Optimization Strategy Analysis

1) Optimization Efficiency Analysis

To analyze the advantages of the RL algorithm in terms of optimization efficiency and accuracy, this paper utilizes the second-order cone relaxation (SOCR) technique to convexify the optimization problem and employs the Gurobi solver to solve the mixed-integer second-order cone programming (MISOCP) problem. The optimization results and computation times for multi-time sections are listed in Table IV.

TABLE IV
SOLUTION INFORMATION COMPARISON

Optimization metrics	Optimization problem types				
	1 h section (MISOCP)	4 h sections (MISOCP)	8 h sections (MISOCP)	24 h sections (MISOCP)	24 h sections (proposed approach)
Daily operating cost (\$)	4650.26	4571.99	4544.66	4467.42	4524.89
Model solution time (s)	29	167	7452	236 736	4
Average memory usage (%)	5	12	37	41	2
Network loss (MW)	0.897	0.863	0.819	0.782	0.788

The comparison of the data in the table indicates that as the coupling factors of multi-time sections increase, the power system can more fully consider the time-varying characteristics of source-load interactions, resulting in a gradual reduction in overall costs. However, the MISOCP approach exhibits limited optimization

performance only when a small number of multi-time sections are considered. With the expansion of the problem scale, the model's computation time increases substantially. In contrast, the decision-making agent constructed through RL is not constrained by the scale of the optimization problem and can provide precise and

complete scheduling actions within seconds, with its optimization cost and network loss only slightly higher than the optimal results of scheduling 24 hour sections.

2) Switch Action Analysis

Table V shows the topology reconstruction strategy of the model within the optimized periods. From this table, it can be seen that the RL model successfully identified four distinct types of topologies suitable for reconstruction to meet system demands at different times. The neural network, by incorporating switch status and system topology into the state and reward functions, effectively captures the relationship between the network structure before and after each period. It operates only a subset of line switches during reconstruction, gradually transforming the system topology into an optimal configuration. This approach substantially alleviates the negative impacts of frequent switch operations on both switching performance and system economic operations.

TABLE V
TOPOLOGY STRATEGY ANALYSIS

Refactoring period	Disconnect branch number	Number of operations
00:00–04:00	36, 37, 6, 35, 13	0
04:00–06:00	36, 27, 6, 10, 13	4
06:00–10:00	36, 37, 6, 35, 13	4
10:00–13:00	36, 37, 6, 35, 34	2
13:00–18:00	36, 37, 6, 35, 13	2
18:00–20:00	36, 37, 6, 35, 34	2
20:00–21:00	36, 27, 6, 35, 13	4
21:00–00:00	36, 37, 6, 35, 13	2

Figure 9 depicts the changes in system network loss after topology reconstruction. Overall, the network loss is lower during the day and increases at night. This is because the PV output is higher during the day and the storage regulation is stronger. In this case, most of the power can be quickly consumed locally, while the load reaches the peak at night. Even though energy storage regulation, the overall load is still high, resulting in a sharp increase in network loss. The algorithm in this paper adjusts the system topology at the right time to reasonably configure the power transmission path, effectively reducing the system network loss in different time scenarios. The average network loss reduction ratio reaches 23.8%, and the minimum network loss reduction ratio is 8.2%.

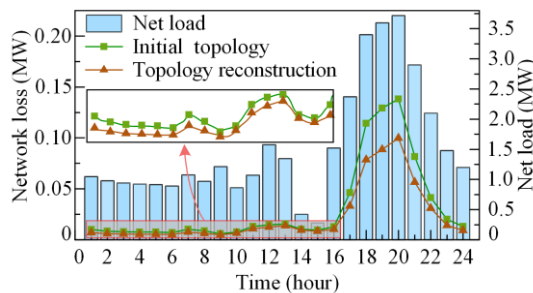


Fig. 9. Comparison of network loss in different topologies.

The voltage deviation problem under high PV penetration is a key issue in distribution network optimization. Figure 10 shows the voltage at each node under

different topologies. Initially, the voltage is concentrated between 0.95 p.u. and 1.02 p.u. After reconstruction, the voltage distribution at each node is more centered around the rated voltage. The voltage fluctuation amplitude during the optimization cycle is significantly reduced, and the local voltage increase during PV output peaks is effectively suppressed, leading to better power quality of the distribution network system. These subtle structural adjustments further enhance the system's ability to absorb PV power.

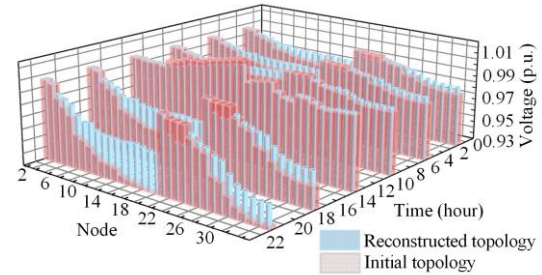


Fig. 10. Comparison of node voltage in different topologies.

3) Energy Storage Action Strategy Analysis

While performing topological reconfiguration, the model provides corresponding charging and discharging strategies for energy storage. As illustrated in Fig. 11, energy storage is charged at night and is discharged during the morning peak hours to achieve peak shaving and valley filling. Meanwhile, it prepares for the expected peak output of photovoltaic generation. The study utilized variance, peak-to-valley difference, and peak-to-valley difference ratio as key optimization metrics to comprehensively evaluate load fluctuations before and after optimization. As detailed in Table VI, the findings indicate that all metrics have undergone notable enhancements. Specifically, the variance metric decreased by about 10%, while the peak-to-valley difference (P-V difference) metric decreased by approximately 6%. The main constraint hindering further advancements in this context is the limitation posed by energy storage capacity and the amplitude of charging and discharging power.

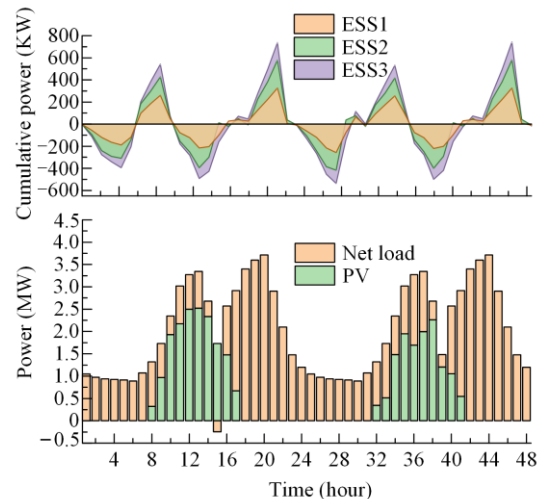


Fig. 11. Load and energy storage strategies.

TABLE VI
COMPARISON OF PAYLOAD EVALUATION INDICATORS

Algorithm	Variance	P-V difference	P-V difference rate (%)
Not optimized	0.951	3.435	92.5
Proposed approach	0.862	3.228	90.7

V. CONCLUSIONS

This paper proposes a decision-making solution framework based on RL, which introduces a topology masking mechanism and an action ensemble screening strategy during the training phase to achieve real-time reconfiguration and operation of distribution network systems integrated with distributed photovoltaic sources. Experimental results indicate that the proposed scheduling strategy can effectively reduce system line losses and stabilize node voltage. Future studies can consider embedding physical constraints directly into neural networks to improve the exploration efficiency and safety of the algorithm.

ACKNOWLEDGMENT

Not applicable.

AUTHORS' CONTRIBUTIONS

Haixiang Zang: conceptualization, methodology, and software. Yongkai Zhao: data curation, software, and writing original draft. Kang Sun: data curation, and writing-reviewing & editing. Guoqiang Sun: software, and writing reviewing & editing. Lilin Cheng, Jingxuan Liu, and Zhinong Wei: writing-reviewing and editing. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (No. U24B2088).

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

DECLARATIONS

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS' INFORMATION

Haixiang Zang received the B.S. degree in electrical engineering, in 2009 from Nanjing Normal University, Nanjing, China, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2014. He is currently a professor with the School of Electrical and Power Engineering, Hohai University, Nanjing. His

research interests include generation of renewable energy and operation and control of power systems.

Yongkai Zhao received the B.S. degree in electrical engineering and automation in 2023 from China University of Petroleum, Qingdao, China. He is currently pursuing the M.S. degree in electrical engineering at Hohai University, Nanjing, China. His research focuses on distribution network optimization based on reinforcement learning.

Kang Sun received the B.S. degree in electrical engineering and automation from the College of Energy and Electrical Engineering, Hohai University, Nanjing, China, in 2019. He is currently pursuing the Ph.D. degree in electrical engineering with the School of Electrical and Power Engineering, Hohai University. From September 2023 to September 2024, he was a visiting scholar at the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. His current research interests include power system state estimation, cyber-physical systems, and high performance computing.

Guoqiang Sun received the B.S., M.S., and Ph.D. degrees in electrical engineering from Hohai University, Nanjing, China, in 2001, 2005, and 2010, respectively. From 2015 to 2016, he was a visiting scholar with North Carolina State University, Raleigh, NC, USA. He is currently a professor with the School of Electrical and Power Engineering, Hohai University. His research interests include power system analysis and economic dispatch and optimal control of integrated energy systems.

Lilin Cheng received the B.S. degree in electrical engineering from Nanjing Normal University, Nanjing, China, in 2017, and the Ph.D. degree in electrical engineering, in 2024 from Hohai University, Nanjing. His research interests include renewable energy integrated power systems and artificial intelligence technologies for smart grids.

Jingxuan Liu received the B.S. degree in electrical engineering and automation in 2021 from Hohai University, Nanjing, China, where he is currently working toward the Ph.D. degree in electrical engineering. He focuses on artificial intelligence technologies for renewable energy integration.

Zhinong Wei received the B.S. degree in electrical engineering from the Hefei University of Technology, Hefei, China, in 1984, the M.S. degree in electrical engineering from Southeast University, Nanjing, China,

in 1987, and the Ph.D. degree in electrical engineering from Hohai University, Nanjing, in 2004. He is currently a professor of electrical engineering with the School of Electrical and Power Engineering, Hohai University. His research interests include power system estimation, integrated energy systems, smart distribution systems, optimization and planning, load forecasting, and integration of distributed generation into electric power systems.

REFERENCES

- [1] W. Zhao, T. Zeng, and Z. Liu *et al.*, "Automatic generation control in a distributed power grid based on multi-step reinforcement learning," *Protection and Control of Modern Power Systems*, vol. 9, no. 4, pp. 39-50, Jul. 2024.
- [2] M. Li, Y. Wang, and P. Peng *et al.*, "Toward efficient smart management: a review of modeling and optimization approaches in electric vehicle-transportation network-grid integration," *Green Energy and Intelligent Transportation*, vol. 3, no. 6, Dec. 2024.
- [3] J. Liu, H. Zang, and T. Ding *et al.*, "A principle-constrained wind field image generation framework for short-term wind power forecasting," *IEEE Transactions on Power Systems*, vol. 40, no. 2, pp. 1790-1801, Mar. 2025.
- [4] Y. Chai, L. Guo, and C. Wang *et al.*, "Hierarchical distributed voltage optimization method for HV and MV distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 968-980, Mar. 2020.
- [5] W. Jiang, Z. Guo, and Y. Pang *et al.*, "Optimal decision-making method for hydrogen-blended integrated energy system based on digital twin models and Stackelberg game theory," *Power System Protection and Control*, vol. 53, no. 11, pp. 72-83, Jun. 2025. (in Chinese).
- [6] Y. Gao, W. Wang, and J. Shi *et al.*, "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357-5369, Nov. 2020.
- [7] Z. Guo, W. Wei, and M. Shahidehpour *et al.*, "Two-timescale dynamic energy and reserve dispatch with wind power and energy storage," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 1, pp. 490-503, Jan. 2023.
- [8] Erdiwansyah, Mahidin, and H. Husin *et al.*, "A critical review of the integration of renewable energy sources with various technologies," *Protection and Control of Modern Power Systems*, vol. 6, no. 1, pp. 1-18, Jan. 2021.
- [9] S. Xu, Q. Liu, and Y. Hu *et al.*, "Decision-making models on perceptual uncertainty with distributional reinforcement learning," *Green Energy and Intelligent Transportation*, vol. 2, no. 2, Apr. 2023.
- [10] N. Yang, X. Li, and Y. Huang *et al.*, "Hierarchical multi-agent deep reinforcement learning for multi-objective dispatching in smart grid," in *Proceedings of 2021 China Automation Congress (CAC)*, Beijing, China, Oct. 2021, pp. 4714-4719.
- [11] X. Sun and J. Qiu, "Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2903-2912, Jul. 2021.
- [12] S. Gao, C. Xiang, and M. Yu *et al.*, "Online optimal power scheduling of a microgrid via imitation learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 861-876, Mar. 2022.
- [13] C. Jiang, Z. Lin, and C. Liu *et al.*, "MADDPG-based active distribution network dynamic reconfiguration with renewable energy," *Protection and Control of Modern Power Systems*, vol. 9, no. 6, pp. 143-155, Nov. 2024.
- [14] T. Han and D. J. Hill, "Learning-based topology optimization of power networks," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1366-1378, Mar. 2023.
- [15] Y. Tao, J. Qiu, and S. Lai *et al.*, "Distributed adaptive robust restoration scheme of cyber-physical active distribution system with voltage control," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 2170-2184, Jan. 2024.
- [16] X. Han, Y. Hao, and Z. Chong *et al.*, "Deep reinforcement learning based autonomous control approach for power system topology optimization," in *Proceedings of 2022 41st Chinese Control Conference (CCC)*, Hefei, China, Jul. 2022, pp. 6041-6046.
- [17] Y. Zheng, Z. Yan, and K. Chen *et al.*, "Vulnerability assessment of deep reinforcement learning models for power system topology optimization," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3613-3623, Jul. 2021.
- [18] R. Si, S. Chen, and J. Zhang *et al.*, "A multi-agent reinforcement learning method for distribution system restoration considering dynamic network reconfiguration," *Applied Energy*, vol. 372, Oct. 2024.
- [19] W. Huang, W. Zheng, and D. J. Hill, "Distribution network reconfiguration for short-term voltage stability enhancement: an efficient deep learning approach," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5385-5395, Nov. 2021.
- [20] J. Qin, Y. Gao, and M. Bragin *et al.*, "An optimization method-assisted ensemble deep reinforcement learning algorithm to solve unit commitment problems," *IEEE Access*, vol. 11, pp. 100125-100136, Sept. 2023.
- [21] B. Wang, H. Zhu, and H. Xu *et al.*, "Distribution network reconfiguration based on NoisyNet deep Q-learning network," *IEEE Access*, vol. 9, pp. 90358-90365, Jun. 2021.
- [22] H. Gao, S. Jiang, and Z. Li *et al.*, "A two-stage multi-agent deep reinforcement learning method for urban distribution network reconfiguration considering switch contribution," *IEEE Transactions on Power Systems*, vol. 39, no. 6, pp. 7064-7076, Nov. 2024.
- [23] Z. Yi, S. Liang, and W. Wang *et al.*, "Power system dispatch: an accelerated safe reinforcement learning approach by incorporating learning from demonstration," *Proceedings of the CSEE*, vol. 44, no. 13, pp.

- 5084-5096, Jul. 2024. (in Chinese)
- [24] C. Li, Y. Xi, and Y. Lu *et al.*, "Resilient outage recovery of a distribution system: co-optimizing mobile power sources with network structure," *Protection and Control of Modern Power Systems*, vol. 7, no. 3, pp. 1-13, Jul. 2022.
- [25] C. Shang, L. Fu, and X. Bao *et al.*, "Dynamic fault reconfiguration of distribution networks in ship power systems based on deep reinforcement learning approach," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 3, pp. 7076-7089, Sept. 2024.
- [26] H. Mao, Z. Zhang, and Z. Xiao *et al.*, "Modelling the dynamic joint policy of teammates with attention multi-agent DDPG," in *Proceedings of Proceedings of the 18th International Conference on Autonomous Agents and Multi Agent Systems*, Montreal, Canada, May 2019, pp. 1108-1116.
- [27] O. E. Egbomwan, S. Liu, and H. Chaoui, "Twin delayed deep deterministic policy gradient (TD3) based virtual inertia control for inverter-interfacing DGs in microgrids," *IEEE Systems Journal*, vol. 17, no. 2, pp. 2122-2132, Jun. 2023.
- [28] Y. Zhou, B. Zhang, and C. Xu *et al.*, "A data-driven method for fast AC optimal power flow solutions via deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1128-1139, Nov. 2020.
- [29] B. Wang, C. Zhang, and Z. Y. Dong, "Interval optimization based coordination of demand response and battery energy storage system considering SOC management in a microgrid," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2922-2931, Oct. 2020.
- [30] L. Zhao, Y. Tang, and Y. Zhang, "Coordinated control of active and reactive power of distribution network with distributed photovoltaic based on scene analysis," in *Proceedings of 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC)*, Beijing, China, Sept. 2019, pp. 753-758.
- [31] X. Zhou, X. Zhang, and H. Zhao *et al.*, "Active disturbance rejection control of a microgrid load-side interface converter based on a DDPG algorithm," *Power System Protection and Control*, vol. 51, no. 21, pp. 66-75, Dec. 2023. (in Chinese)