Active Distribution Network Voltage Control Strategy Considering the Emergency Faults: A Knowledge Fusion-Enhanced Multi-Agent Meta-Reinforcement Learning Method

Haopeng An
School of Mechanical and Electrical Engineering
University of Electric Science and Technology of China
Chengdu, China
haopengan@std.uestc.edu.cn

Guangdou Zhang
Department of Electrical Engineering
Tsinghua University
Beijing, China
gdzhang@mail.tsinghua.edu.cn

Jian Li*
School of Mechanical and Electrical Engineering
University of Electric Science and Technology of China
Chengdu, China
leejian@uestc.edu.cn

Yalong Mai
School of Mechanical and Electrical Engineering
University of Electric Science and Technology of China
Chengdu, China
ylmai@std.uestc.edu.cn

***Abstract*—The increasing penetration of renewable energy has facilitated the transition from traditional distribution networks to active distribution networks (ADN). However, the integration of renewable energy exacerbates the complexity and fragility of ADN, leading to more frequent fault occurrence. Maintaining voltage within tolerance during the fault event is critical for ADN. This paper proposes a knowledge fusion-enhanced multi-agent meta-reinforcement learning voltage control strategy for photovoltaic (PV) integrated ADN. The proposed strategy is applied to batches of pre-training tasks for knowledge extraction. An average cosine similarity-based masking mechanism (ACSMM) is embedded into the meta-learning process to enhance the effect of knowledge fusion from various pre-training tasks. This enables the agents to rapidly fine-tune and make optimal decisions based solely on limited data during unpredictable emergency failures. Case studies validate the effectiveness of the proposed strategy in controlling the voltage into safety domain.**

***Keywords*—active distribution network, meta-learning, reinforcement learning, renewable energy**

I. INTRODUCTION

As the penetration rate of renewable energy continues to rise, active distribution networks (ADN) with local consumption capabilities of renewable energy have seen rapid development in recent years [1]. The deployment of numerous advanced measurement and monitoring devices has enhanced the efficiency of the ADN and its capacity to integrate renewable energy. However, the integration of numerous metering devices and renewable energy sources increases the risk of emergency fault events occurring, thereby heightening the risk of voltage violations in ADN [2] [3].

Emergency events in distribution networks are characterized by high uncertainty. Various model-based control methods have been proposed to alleviate the voltage violations under uncertain emergency faults. A robust voltage control strategy in [4] is proposed for the V2G procedure under ADN measurement anomalies. A H_∞ controller is constructed in [5] to alleviate the threat of cyber-attack. Nazir et al. [6] proposed a robust optimization theory-based voltage control algorithm for active distribution networks considering the uncertainty of renewable energy output. However, most existing robust control strategies are designed based on the worst-case scenario, leading to insufficient flexibility.

The high level of uncertainty inherent in emergency faults can be mitigated with the help of advanced machine learning optimization methods. Among the various machine learning methods, reinforcement learning (RL) methods have been successfully applied to ADN voltage control due to their high adaptability without the need for accurate environment modeling. A multi-agent deep reinforcement learning strategy is constructed in [7] for PV-integrated ADN voltage control. It is noted that the unpredictable emergency faults usually happen in renewable energy integrated ADN, leading to RL-based agents needing to be retrained, which is time-consuming and requires a lot of operational data. Since emergency events demand fast responses and only limited data is available under emergency faults, most reinforcement learning–based approaches are unable to satisfy this requirement. In recent years, meta-reinforcement learning has been introduced into the power system domain to train learners to achieve rapid adaptation [8]. However, due to variations in fault locations, load conditions, and network topologies, emergency scenarios may exhibit significant differences. In such contexts, traditional meta-learning may encounter gradient conflicts [9], thereby reducing its generalization. Consequently, the meta-reinforcement learning for voltage control in ADNs considering emergency conditions remains in its infancy and needs further exploration.

To this end, this paper proposes a knowledge fusion-enhanced multi-agent meta-reinforcement learning method for ADN voltage control considering the unpredictable emergency faults. The main contributions are summarized as follows:

1) The proposed meta-learning–based approach enables fast and effective inverter output decision-making under new emergency events with only a small amount of fine-tuning data. Compared with existing RL approaches that require large-

This work is supported by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (No. GZC20250341 and 2025M780469.) and National Natural Science Foundation of China (NSFC, No. 52407001) and Yibin City Dual City-University Agreement Special Scientific Research Funding Project in YiBin, Sichuan,China (No. YBSCXY2024010003).

scale data for training, it exhibits advantages in rapid deployment.

2) An average cosine similarity-based masking mechanism (ACSMM) is embedded into the outer loop of the meta-learning process to enhance the effect of knowledge fusion from various pre-training tasks. This is achieved by mitigating gradient conflicts through scaling the learning rates across different layers of the agents.

3) The proposed method adopts a centralized training and decentralized execution (CTDE) framework. During the meta-training stage, agents with the multi-agent proximal policy optimization (MAPPO) algorithm fully utilize global information, whereas in the execution phase, they make real-time independent decisions based on local observations. This framework enhances the flexibility of the system.

The remainder of this paper is organized as follows: Section II provides the problem formulation. The proposed methodology is formulated in Section III. Case studies are conducted in Section IV. Section V summarizes this paper.

II. Problem Formulation

The goal of voltage control is to minimize the voltage deviation in ADN. Such minimization is realized by controlling the reactive power outputs of PV. In this context, the objective function is formulated as follows:

$$\min_{Q_i^{PV}} \sum_{i\in N} \left|V_{i,t} - V_0\right| + \beta \sum_{i\in N} P_{i,t}^{loss} \tag{1}$$

Subject to

$$\begin{aligned} &P_{i,t}^{PV} - P_{i,t}^{L} = V_{i,t}^2 \sum G_{ij,t} \\ &-V_{i,t} \sum V_{j,t} \left(G_{ij,t} \cos\theta_{ij,t} + B_{ij,t} \sin\theta_{ij,t}\right) \\ &Q_{i,t}^{PV} - Q_{i,t}^{L} = -V_{i,t}^2 \sum G_{ij,t} \\ &+V_{i,t} \sum V_{j,t} \left(G_{ij,t} \sin\theta_{ij,t} + B_{ij,t} \cos\theta_{ij,t}\right) \end{aligned} \tag{2}$$

$$P_{i,t}^{loss} = \frac{\left(P_{i,t}^{L} - P_{i,t}^{PV}\right)^2 + \left(Q_{i,t}^{L} - Q_{i,t}^{PV}\right)^2}{V_{i,t}^2} \tag{3}$$

$$\left|Q_{i,t}^{PV}\right| \le Q_{i,t,\max}^{PV}, i \in N^{PV} \tag{4}$$

$$(Q_{i,\max}^{PV})^2 + (P_{i,t}^{PV})^2 = (S_i^{PV})^2, \forall i \in N^{PV} \tag{5}$$

$$0.95 \text{ p.u.} \le V_{i,t} < 1.05 \text{p.u.}, i \in N/0 \tag{6}$$

$$V_0 = 1.0 \text{p.u.} \tag{7}$$

where $V_{i,t}$ is the voltage of i-th node and V_0 represents the voltage of reference node. Eq. (2) constrains the active and reactive power balance of ADN where $P_{i,t}^{PV}/Q_{i,t}^{PV}$ is the active/reactive power output of PV of node i at time t. $G_{ij,t}/B_{ij,t}$ represent the susceptance and conductance of branch between node i and j at time t. θ_{ij} represent the phase angle difference between de i and j at time t. $P_{i,t}^{L}/Q_{i,t}^{L}$ are the active/reactive power load of node i at time t. Eq. (3) provides the calculation of the load shedding power of node i. Eq. (4)-(5) constrain the upper limit of reactive power output of PV of node i at time t. Voltage constraints of each node are formulated in Eq. (6)-(7) to ensure the safety and optimal operation of the ADN.

III. Methodology

To explore the voltage control method of ADN considering the unpredictable emergency faults, a knowledge fusion-enhanced multi-agent meta-reinforcement learning method is proposed. The optimization model is reconstructed as a Markov game (MG), where the core elements include agents, states, and actions, and the reward functions are constructed based on the optimization model. A meta-learning framework with a double loop is constructed to enable agents to rapidly converge in the face of unseen new tasks via the acquisition of prior knowledge. In the inner loop, a multi-agent deep reinforcement learning approach with the CTDE framework is employed for rapid adaptation to a specific task. The objective of outer loop meta-learning is the optimization of meta-parameters such that all tasks are rapidly adapted in the inner loop. This is achieved by training agents in multiple batches of tasks to adjust the initial parameters to improve cross-task knowledge extraction. It is noted that differences between pre-training tasks can lead to gradient conflicts in the meta-learning update process, thereby reducing the effectiveness of knowledge fusion. To enhance the effect of knowledge fusion in the outer loop, an ACSMM method is constructed. The proposed ACSMM aims to identify layers of agents that experience gradient conflicts across different tasks and slow down updates in those layers to mitigate such conflicts.

A. The Construction of a Markov Game

To improve operational efficiency, the large-scale ADN is divided into multiple sub-regions, each equipped with several photovoltaic (PV) modules. Each module is connected to an inverter that regulates reactive power to maintain the voltage within the tolerance value. In this framework, each region can be regarded as an "agent," and the voltage control of ADN is formulated as a Markov game (MG) as follows:

Agent: In this MG, each sub-region is represented as an agent.

State: The global state set at time t, S_t, includes all agents' observations. For agent j, the observation contains active/reactive power load of all nodes in its corresponding sub-region and active/reactive power of PVs in its corresponding sub-region, which is formulated as follows:

$$o_t^j = (\boldsymbol{P}_{j,t}^{PV}, \boldsymbol{Q}_{j,t}^{PV}, \boldsymbol{P}_{j,t}^{L}, \boldsymbol{Q}_{j,t}^{L}) \tag{8}$$

Action: The action set at time t, A_t, includes all agents' actions. For agent j, the action a_t^j contains the control variables within the j-th sub-region. The continuous action is represented as $a_t^j = \alpha_{j,t} \in \{-1,1\}$, indicating the proportion of maximum reactive power generated by the inverter. Hence, the control variable of agent j is formulated as follows:

$$Q_{j,t}^{PV} = \alpha_{j,t} \sqrt{(S_{j,t}^{PV})^2 - (P_{j,t}^{PV})^2} \tag{9}$$

where $S_{j,t}^{PV}$ represent the capacity of the PV at node j.

State Transition: The transition probability function is applied to describe the dynamic characteristics of the environment as follows:

$$p\left(S' \mid S, A\right) = \text{Prob}\left(S_t = S' \mid S_t = S, A_{t-1} = A\right) \tag{10}$$

It can be observed from the above equation that the transition probability of the current state S_t is related only with the previous state S_{t-1} and actions A_{t-1}.

Reward: The reward function contains two parts: voltage violation penalty and active power load shedding penalty. The reward function is formulated as follows:

$$r_t = \sum_{i\in N} r_{i,t}^{voltage} + \beta \sum_{i\in N} P_{i,t}^{loss} \tag{11}$$

$$r_{i,t}^{voltage} = \begin{cases} -\left|1-V_{i,t}\right| & V_{i,t} \in (0.98,1.02) \\ -2\left|1-V_{i,t}\right| & V_{i,t} \in [0.95,0.98) \cup V_{i,t} \in (1.02,1.05] \\ -200 & \text{otherwise} \end{cases} \tag{12}$$

where $V_{i,t}$ and $P_{i,t}^{loss}$ is calculated by Eq. (2)-(5). $r_{i,t}^{\text{voltage}}$ represents the voltage violation penalty, which is formulated by the piecewise function in Eq. (12). The piecewise reward function enables the agent to distinguish voltage states with different severity levels, thus improving control stability and accelerating convergence during voltage regulation.

B. Solution via MAPPO Algorithm

The MG problem can be solved by modeling each sub-region as a proximal policy optimization (PPO) agent within the centralized training, decentralized execution (CTDE) framework. In the PPO strategy, each agent consists of an actor network and a critic network, both of which are composed of deep neural networks (DNNs). The actor network takes the state from Eq. (8) as input and outputs the action in Eq. (9), while the critic network outputs an evaluation of the actor's decision. In the decentralized execution stage, each actor network receives local observations and outputs real-time control decisions. Owing to the critic network's explicit modeling of other agents' policies during pretraining, actors can learn coordinated behaviors. As a result, cooperation is able to emerge from local observations alone, greatly reducing communication overhead and enhancing scalability to large-scale systems.

The actor networks are trained by maximizing their objectives as follows:

$$L_i(\theta_i^u) = \max \sum_{\tau} r_i(\tau_i)\, p_{\theta_i^u}(\tau_i) \tag{13}$$

where θ_i^u represents the in i-th agent's actor network parameters. $r_i(\tau)$ is the reward. $\tau_i = \{S_1, A_1, S_2, A_2, \ldots, S_t, A_t\}$ denotes the trajectory collected by the agents. $p_{\theta_i^u}(\tau)$ indicates the distribution of samples. By introducing the clip function and generalized advantage estimation, Eq. (13) can be written as follows:

$$L_i^{\text{CLIP}}\left(\theta_i^u\right) = \sum_{(o_{i,t},a_{i,t})} \min \begin{pmatrix} \dfrac{P_{\theta_i^u}\left(a_{i,t}\middle|\, o_{i,t}\right)}{p_{\theta_i^u}\left(a_{i,t}\middle|\, o_{i,t}\right)} \cdot E_i\left(S_t\right), \\ \text{clip}\left(\dfrac{P_{\theta_i^u}\left(a_{i,t}\middle|\, o_{i,t}\right)}{p_{\theta_i^u}\left(a_{i,t}\middle|\, o_{i,t}\right)}, 1-\zeta, 1+\zeta\right) \cdot E_i\left(S_t\right) \end{pmatrix} \tag{14}$$

where clip$(\cdot)$ represents the clip function and $E_i(\cdot)$ is the generalized advantage estimation.

Given the above optimizations, the update of actor network parameters is formulated as follows:

$$\theta_i^u \leftarrow \theta_i^u + \eta_u \nabla_{\theta_i^u} J^{\text{CLIP}}\left(\theta_i^u\right) \tag{15}$$

where η_u is the learning rate of the actor network. The critic networks are trained according to global states and updated via a loss function to improve the accuracy of the estimate:

$$L\left(\theta_i^Q\right) = \mathrm{E}_{\mu'}\left[r_{i,t} + \gamma V^{\pi}\left(S_{t+1}\right) - V^{\pi}\left(S_t\right)^2\right] \tag{16}$$

$$\theta_i^Q \leftarrow \theta_i^Q + \eta_Q \nabla_{\theta_i^Q} L\left(\theta_i^Q\right) \tag{17}$$

where $V^{\pi}(\cdot)$ represents the scalar value of the function output by critic network. θ_i^Q is the parameters of i-th critic network and η_Q is the learning rate of i-th critic network.

C. Meta-learning with ACSMM Embedded

The meta-learning aims to adjust the initial parameters of agents to appropriate values, enabling rapid deployment with minimal gradient update steps described by Eq. (15) and Eq. (17) when confronting unpredictable emergencies. This is achieved by pre-training and updating initial parameters at diverse batch tasks that encompass various emergency failure scenarios. These initial parameters are also referred to as meta-parameters. The traditional update procedure of actor network meta-parameters and critic network meta-parameters are formulated as follows:

$$\theta_i^{\text{meta},u} \leftarrow \theta_i^{\text{meta},u} + \varepsilon_{\text{meta}}^u \frac{1}{m} \sum_{h=1}^{m} \left(\theta_{i,h}^u - \theta_i^{\text{meta},u}\right) \tag{18}$$

$$\theta_i^{\text{meta},Q} \leftarrow \theta_i^{\text{meta},Q} + \varepsilon_{\text{meta}}^Q \frac{1}{m} \sum_{h=1}^{m} \left(\theta_{i,h}^Q - \theta_i^{\text{meta},Q}\right) \tag{19}$$

where $\varepsilon_{meta}^u / \varepsilon_{meta}^Q$ denotes the meta-learning rate of actor network/critic network. $\theta_{i,h}^u / \theta_{i,h}^Q$ represents the parameters of actor network/critic network obtained from h-th tasks. m is the number of tasks in the batch

It is noted that, in the pretraining stage, task discrepancies can cause gradient conflicts that hinder effective knowledge integration during meta-parameter updates [10]. To address this issue, ACSMM aims to compute the cosine similarity to detect conflicting gradients in the agents and alleviate the gradient conflict by reducing the learning rate, thereby improving the knowledge fusion effect.

To illustrate the procedure of the proposed ACSMM, actor networks and critic networks of the i-th agents are written as follows:

$$\theta_i^u = \begin{pmatrix} \theta_i^{u,1} & \theta_i^{u,2} & \cdots & \theta_i^{u,p} \end{pmatrix} \tag{20}$$

$$\theta_i^Q = \begin{pmatrix} \theta_i^{Q,1} & \theta_i^{Q,2} & \cdots & \theta_i^{Q,q} \end{pmatrix} \tag{21}$$

where $\theta_i^{u,p} / \theta_i^{Q,q}$ represents the weight and bias of p-th/q-th layers of actor network/critic network.

In this way, the meta-parameter variations for the i-th actor network/critic network across different tasks within the same batch can be expressed respectively as follows:

$$\begin{pmatrix} g_{i,1}^u \\ g_{i,2}^u \\ \vdots \\ g_{i,m}^u \end{pmatrix} = \begin{pmatrix} \theta_{i,1}^{u,1} - \theta_i^{\text{meta},u,1} & \theta_{i,1}^{u,2} - \theta_i^{\text{meta},u,2} & \cdots & \theta_{i,1}^{u,p} - \theta_i^{\text{meta},u,p} \\ \theta_{i,2}^{u,1} - \theta_i^{\text{meta},u,1} & \theta_{i,2}^{u,2} - \theta_i^{\text{meta},u,2} & \cdots & \theta_{i,1}^{u,p} - \theta_i^{\text{meta},u,p} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{i,m}^{u,1} - \theta_i^{\text{meta},u,1} & \theta_{i,m}^{u,2} - \theta_i^{\text{meta},u,2} & \cdots & \theta_{i,m}^{u,p} - \theta_i^{\text{meta},u,p} \end{pmatrix} \tag{22}$$

$$\begin{pmatrix} g_{i,1}^{Q} \\ g_{i,2}^{Q} \\ \vdots \\ g_{i,m}^{Q} \end{pmatrix} = \begin{pmatrix} \theta_{i,1}^{Q,1} - \theta_{i}^{\text{meta},Q,1} & \theta_{i,1}^{Q,2} - \theta_{i}^{\text{meta},Q,2} & \cdots & \theta_{i,1}^{Q,p} - \theta_{i}^{\text{meta},Q,q} \\ \theta_{i,2}^{Q,1} - \theta_{i}^{\text{meta},Q,1} & \theta_{i,2}^{Q,2} - \theta_{i}^{\text{meta},Q,2} & \cdots & \theta_{i,1}^{Q,p} - \theta_{i}^{\text{meta},Q,q} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{i,m}^{Q,1} - \theta_{i}^{\text{meta},Q,1} & \theta_{i,m}^{Q,2} - \theta_{i}^{\text{meta},Q,2} & \cdots & \theta_{i,m}^{Q,p} - \theta_{i}^{\text{meta},Q,q} \end{pmatrix} \quad (23)$$

For the j-th layer of the i-th actor network/critic network, the average cosine similarity in the same batch is formulated as follows:

$$ACS_{i,j}^{u} = \frac{1}{m(m-1)} \sum_{1 \le s \ne t \le N} \frac{g_{i,s}^{u}(j) \cdot g_{i,t}^{u}(j)}{\left\| g_{i,s}^{u}(j) \right\| \left\| g_{i,t}^{u}(j) \right\|}, \forall j \in \{1,2,...,p\} \quad (24)$$

$$ACS_{i,j}^{Q} = \frac{1}{m(m-1)} \sum_{1 \le s \ne t \le N} \frac{g_{s}^{Q}(j) \cdot g_{t}^{Q}(j)}{\left\| g_{s}^{Q}(j) \right\| \left\| g_{t}^{Q}(j) \right\|}, \forall j \in \{1,2,...,q\} \quad (25)$$

where $ACS_{i,j}^{u}/ACS_{i,j}^{Q}$ represent the average cosine similarity of j-th layer of actor network/critic network at i-th agent. $ACS_{i,j}^{u}/ACS_{i,j}^{Q}$ reflect the degree of gradient conflict across all tasks within the same batch for j-th layer of actor network/critic network at i-th agent. A smaller a $ACS_{i,j}^{u}/ACS_{i,j}^{Q}$ indicates more severe conflict, and the meta-updating learning rate for that layer should be reduced to mitigate gradient conflict.

The SoftMax function is used to calculate the scalar value of the j-th layer of the actor network/critic network at the i-th agent for meta-learning rate adjustment as follows:

$$\gamma_{i,j}^{u} = p \cdot \text{softmax}(ACS_{i,j}^{u}), \forall j \in \{1,2,...,p\} \quad (26)$$

$$\gamma_{i,j}^{Q} = q \cdot \text{softmax}(ACS_{i,j}^{Q}), \forall j \in \{1,2,...,q\} \quad (27)$$

The meta-learning rate with ACSMM can be written as follows:

$$\varepsilon_{i,\text{meta}}^{u,n} = \varepsilon_{\text{meta}}^{u} \cdot \begin{pmatrix} \gamma_{i,1}^{u} & \gamma_{i,2}^{u} & \cdots & \gamma_{i,p}^{u} \end{pmatrix} \quad (28)$$

$$\varepsilon_{i,\text{meta}}^{Q,n} = \varepsilon_{\text{meta}}^{Q} \cdot \begin{pmatrix} \gamma_{i,1}^{Q} & \gamma_{i,2}^{Q} & \cdots & \gamma_{i,q}^{Q} \end{pmatrix} \quad (29)$$

where n represents the number of batches. In this way, the meta parameter updates procedure with ACSMM can be expressed by rewriting Eqs. (18)-(19) as follows:

$$\theta_{i}^{\text{meta},u} \leftarrow \theta_{i}^{\text{meta},u} + \varepsilon_{i,\text{meta}}^{u,n} \sum_{h=1}^{m} \left(\theta_{i,h}^{u} - \theta_{i}^{\text{meta},u} \right) \quad (30)$$

$$\theta_{i}^{\text{meta},Q} \leftarrow \theta_{i}^{\text{meta},Q} + \varepsilon_{i,\text{meta}}^{Q,n} \sum_{h=1}^{m} \left(\theta_{i,h}^{Q} - \theta_{i}^{\text{meta},Q} \right) \quad (31)$$

Eq. (30)-(31) adjusts the meta-learning rate according to the average cosine similarity to mitigate gradient conflicts and enhance knowledge fusion. In particular, a higher average cosine similarity indicates smaller gradient conflicts, in which case the corresponding layer should be updated normally. Conversely, when the average cosine similarity is low, the meta-learning rate is reduced to alleviate gradient conflicts. The flowchart of the proposed method is shown in Fig. 1.

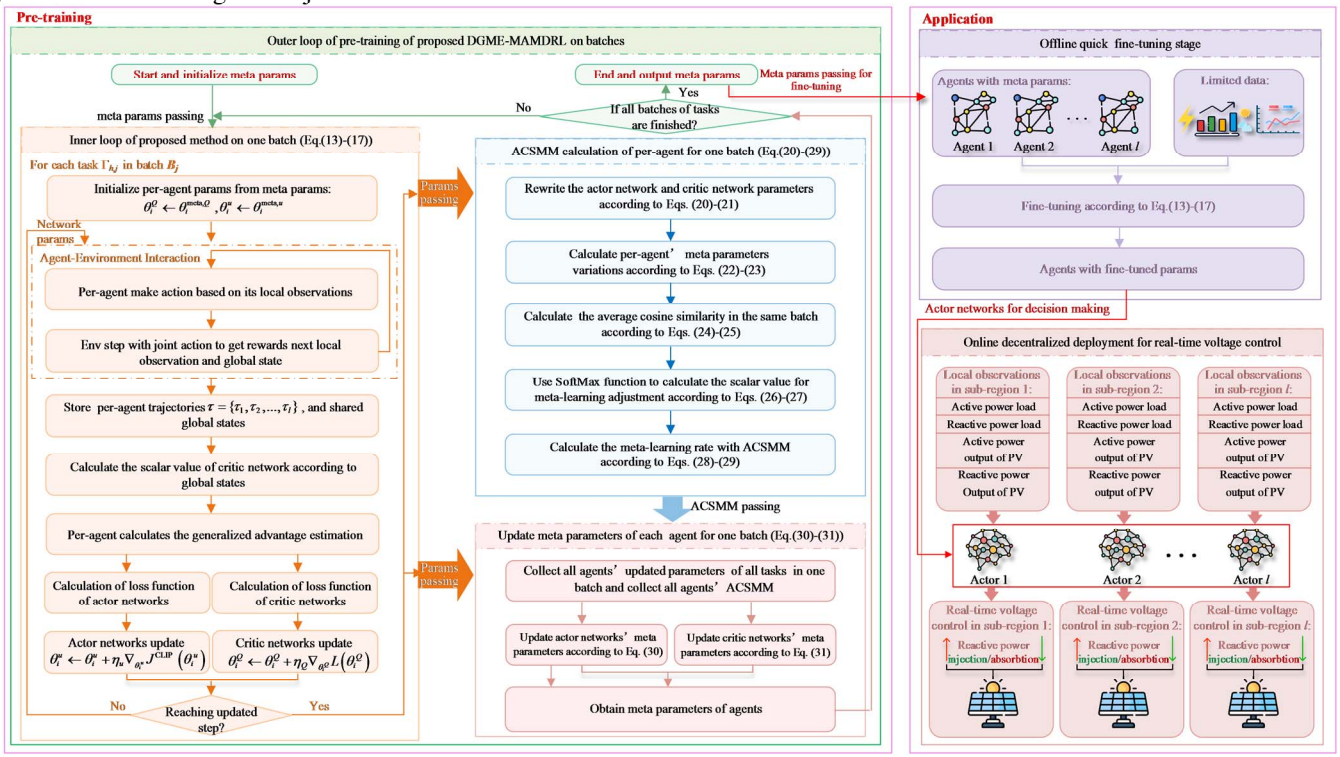

Fig. 1. Flowchart of the proposed voltage control strategy

IV. Case Study

Case studies are conducted on the IEEE 33-bus system [11] with 4 PVs equipped to validate the effect of the proposed method. Gaussian noise is added to the load data to ensure realism. Solar data is obtained from the Belgian grid operator Elia Group with 3-min resolution to align with the AND's real-time control procedure. The technical parameters are provided in TABLE I

The proposed method is trained based on the algorithm in Fig. 1. The rewards and relative improvement ratios are illustrated in Fig. 2. It can be observed from Fig. 2 (a) that the proposed method converges after approximately 300 episodes, with the reward value at convergence ranging from -4 to 2.

This indicates that the proposed method exhibits strong stability. Fig 2 (b) presents ablation experiments, with MAPPO serving as the baseline. As shown in Fig 2 (b), compared to the MAPPO method, the meta-learning approach without ACSMM achieves about 110% performance improvement after 20 adaptations. Incorporating the ACSMM mechanism further enhances performance by 120%. These results demonstrate the effectiveness of both the meta-learning approach and the proposed ACSMM method in enhancing the agent's fast adaptation capabilities and knowledge fusion abilities.

Besides, a stem plot is constructed to comprehensively illustrate the variation of the proposed method's reward with adaptation steps under 50 test scenarios. As evidenced in Fig. 3, the proposed method consistently achieves high reward values after 15 adaptation iterations across all test tasks, demonstrating its superior cross-task adaptability.

TABLE I
TECHNICAL PARAMETERS OF THE PROPOSED STRATEGY

Parameters	Value
Minibatch size	100
ξ	0.1
η_u/η_Q	5e-2/5e-2
$\varepsilon^u_{meta}/\varepsilon^Q_{meta}$	5e-3

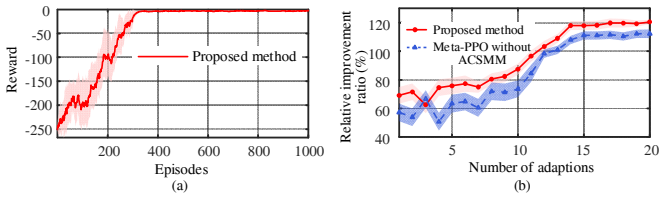

Fig. 2. (a) Reward over 1000 episodes of the proposed method on the 33-node system. (b) Relative improvement ratio curves for different methods on the 33-node system.

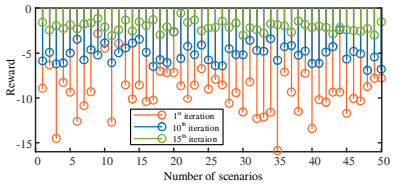

Fig. 3. Stem plot of reward for different test tasks on a 33-node system.

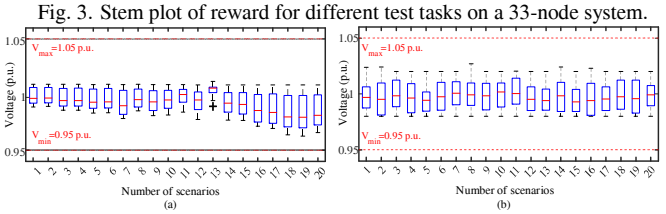

Fig.4 (a). voltage distribution under 20 scenarios on 33-node system. (b). voltage distribution under 20 scenarios on 69-node system.

The voltage distributions under 20 test scenarios on both the IEEE 33-node system and 69-node system are shown in Fig. 4. It can be observed that the voltage distribution is centered around 1 p.u. and remains within the range of 1.05 p.u. to 0.95 p.u. in all scenarios, meeting safety requirements.

V. CONCLUSION

This paper proposes a knowledge fusion-enhanced multi-agent meta-reinforcement learning method for active distribution network voltage control considering the emergency faults. The proposed method aims to maintain voltage within permissible limits with minimal data fine-tuning when encountering unpredictable faults. This is achieved by acquiring prior knowledge through multi-batch task pre-training using the constructed meta-reinforcement learning strategy. An average cosine similarity masking mechanism is proposed to be embedded within the meta-learning process to enhance knowledge fusion effectiveness. Case studies conducted on the IEEE 33-node distribution network and the IEEE 69-node distribution network validate that the proposed method achieved fast convergence when facing new emergency faults and realized voltage control within the tolerance. Future work will focus on safety constrained RL method to enhance the safety of control action.

REFERENCES

[1] G. Varathan and B. E. J, "Review article: A review of uncertainty management approaches for active distribution system planning," *Renew. Sustain. Energy Rev.*, vol. 205, no. August, p. 114808, 2024.

[2] I. Hamdeen, E. A. Badran, M. F. Kotb, and M. Elgamal, "Self-healing multi-agent techniques in electric power distribution systems : A review," *Renew. Sustain. Energy Rev.*, vol. 224, no. November 2024, p. 116132, 2025.

[3] W. Gao, R. Fan, Q. Huang, Y. Li, and D. W. Gao, "A Meta-Strategy Framework With Deep Reinforcement Learning for Damping Interarea Oscillation via High-Voltage DC Transmission," *IEEE Trans. Ind. Informatics*, vol. PP, pp. 1–11, 2025.

[4] H. An *et al.*, "A Robust V2G Voltage Control Scheme for Distribution Networks Against Cyber Attacks and Customer Interruptions," *IEEE Trans. Smart Grid*, vol. 15, no. 4, pp. 3966–3978, 2024.

[5] G. Zhang, J. Li, S. Member, O. Bamisile, Y. Xing, and D. Cai, "Load Frequency Control Scheme for Multi-Area Power System Under Cyber-Attacks and Time-Varying Delays," vol. 8950, no. c, 2022.

[6] F. U. Nazir, R. A. Jabr, and I. I. Parameters, "Affinely Adjustable Robust Volt / VAr Control Without Centralized Computations," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 656–667, 2023.

[7] J. Huang, H. Zhang, D. Tian, Z. Zhang, C. Yu, and G. P. Hancke, "Engineering Applications of Artificial Intelligence Multi-agent deep reinforcement learning with enhanced collaboration for distribution network voltage control," *Eng. Appl. Artif. Intell.*, vol. 134, no. August 2023, p. 108677, 2024.

[8] G. Zhang *et al.*, "Meta-learning and proximal policy optimization driven two-stage emergency allocation strategy for multi-energy system against typhoon disasters," *Renew. Energy*, vol. 237, no. PC, p. 121806, 2024.

[9] K. M. Lee, "Learning to Forget for Meta-Learning," pp. 2376–2384, 2020, doi: 10.1109/CVPR42600.2020.00245.

[10] F. Chen, J. Yan, Y. Liu, Y. Yan, and L. B. Tjernberg, "A novel meta-learning approach for few-shot short-term wind power forecasting," *Appl. Energy*, vol. 362, no. February, p. 122838, 2024.

[11] S. H. Dolatabadi, M. Ghorbanian, P. Siano, S. Member, N. D. Hatziargyriou, and L. Fellow, "An Enhanced IEEE 33 Bus Benchmark Test System for Distribution System Studies," vol. 36, no. 3, pp. 2565–2572, 2021.