

# Stabilizing Voltage in Power Distribution Networks via Multi-Agent Reinforcement Learning with Transformer

Minrui Wang\*

Mingxiao Feng\*

wangminrui0804@mail.ustc.edu.cn  
fmxustc@mail.ustc.edu.cn  
University of Science and Technology  
of China  
Hefei, Anhui, China

Wengang Zhou†

zhwg@ustc.edu.cn

University of Science and Technology

of China

Hefei, Anhui, China

Institute of Artificial Intelligence,  
Hefei Comprehensive National  
Science Center  
Hefei, Anhui, China

Houqiang Li†

lihq@ustc.edu.cn

University of Science and Technology

of China

Hefei, Anhui, China

Institute of Artificial Intelligence,  
Hefei Comprehensive National  
Science Center  
Hefei, Anhui, China

## ABSTRACT

The increased integration of renewable energy poses a slew of technical challenges for the operation of power distribution networks. Among them, voltage fluctuations caused by the instability of renewable energy are receiving increasing attention. Utilizing MARL algorithms to coordinate multiple control units in the grid, which is able to handle rapid changes of power systems, has been widely studied in active voltage control task recently. However, existing approaches based on MARL ignore the unique nature of the grid and achieve limited performance. In this paper, we introduce the transformer architecture to extract representations adapting to power network problems and propose a Transformer-based Multi-Agent Actor-Critic framework (T-MAAC) to stabilize voltage in power distribution networks. In addition, we adopt a novel auxiliary-task training process tailored to the voltage control task, which improves the sample efficiency and facilitates the representation learning of the transformer-based model. We couple T-MAAC with different multi-agent actor-critic algorithms, and the consistent improvements on the active voltage control task demonstrate the effectiveness of the proposed method.<sup>1</sup>

## CCS CONCEPTS

- Computing methodologies → Multi-agent reinforcement learning.

## KEYWORDS

Multi-agent Reinforcement Learning, Active Voltage Control, Transformer

\*Both authors contributed equally to this research.

†Corresponding authors: Wengang Zhou and Houqiang Li.

<sup>1</sup>Code will be released at <https://github.com/cjdr/T-MAAC>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

KDD '22, August 14–18, 2022, Washington, DC, USA.

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9385-0/22/08...\$15.00

<https://doi.org/10.1145/3534678.3539480>

## ACM Reference Format:

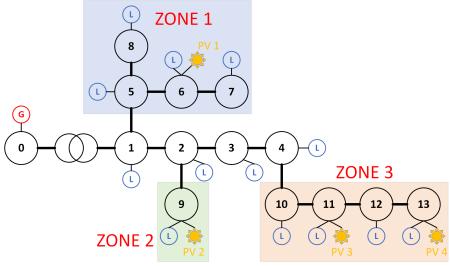
Minrui Wang, Mingxiao Feng, Wengang Zhou, and Houqiang Li. 2022. Stabilizing Voltage in Power Distribution Networks via Multi-Agent Reinforcement Learning with Transformer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3534678.3539480>

## 1 INTRODUCTION

The development and utilization of renewable energy is critical for addressing current energy and environmental concerns. In recent years, distributed generations (DGs), e.g., roof-top photovoltaics (PVs), have been steadily connected to power distribution networks because of its particular environmental friendliness, economy and flexibility. However, the increasing penetration of PVs in the distribution network may cause voltage swing due to their rapid active power changes. The voltage fluctuation can be alleviated by utilizing the control flexibility provided by PV inverters and other controllable devices such as static var compensators (SVCs)[16]. Therefore, an elaborate scheme is required to coordinate the control between these distributed devices based on local information to ensure stable operation of the entire power system, which is called active voltage control[26].

There have been some previous efforts dedicated to active voltage control, which can be classified into three broad groups from the perspective of the control framework: centralized, distributed autonomous, and distributed cooperative control[4]. As a promising solution, the distributed cooperative control enables collaboration between distinct control units via limited communication links. Among them, some approaches[3–5, 16, 27] apply multi-agent reinforcement learning (MARL) to active voltage control. These MARL algorithms are based on the centralized training and decentralized execution framework, which extract knowledge from historical data and simulation environment. The learned strategies can be deployed to the grid and achieve cooperative control without any communication devices. These attempts of applying MARL to power network tasks have attracted a lot of attention because of their strong adaptability to the unknown dynamic in real-time.

For active voltage control problem, distributed control units are treated as agents, observing information about nodes in a zone of the grid. Figure 1 shows an example, in which the whole grid



**Figure 1: An example in power distribution network.** Each bus in the distribution network is considered as a node in the graph. The whole network are divided into 4 zones connected to the main branch (node 1-4). We need to control voltages on node 1-13 and node 0 represent the main grid with constant voltage. G denotes an external generator; L denotes load connected to the node; and the sun emoji denotes the location a PV installed in.

is divided into three zones and the sun emoji represents the location where a PV is installed. Each PV inverter installed in a PV is regarded as a agent and observes information of nodes in the corresponding zone. In prior works[3–5, 16, 27], the MLP-based policy networks and MLP-based critic network are applied to parameterize the policy functions and action-value functions, respectively[1, 10, 12, 19, 30]. Besides, additional paddings are appended to each observation to guarantee that all observations have the same dimension. After that, the padded observations are mapped to control actions via policy networks whose parameters are shared among all agents.

It is worth noting that directly applying the above routine settings in the MARL community to the active voltage control task encounters a number of challenges:

- (1) **Inconsistent number of nodes observed by various agents.** For example, as shown in Figure 1, there are 4 nodes in zone 1 and 3, but only 1 node in zone 2.
- (2) **Inconsistent topology of nodes in each observation.** In Figure 1, both zone 1 and zone 3 include 4 nodes, however they are connected in quite different ways.
- (3) **Inconsistent importance of nodes in a zone.** For example, typically, node 6 installed with PV has more frequent voltage fluctuation than other nodes in zone 1, which means that the PV 1 must take more attention to node 6 when making decisions. Current methods ignore these challenges and simply concatenate nodes information directly to obtain observations, with an assumption that neural networks are capable of automatically modeling the relationship between nodes in a sub-grid and decoupling observations smartly to address above challenges. By following these settings, these approaches handle all information received from different agents in a same way and treat nodes in a observation uniformly with no regard of the topology of these nodes in the zone, which leads to limited representation learning capability.

To address the above challenges, we propose a transformer [25] based multi agent actor-critic framework (abbreviation as T-MAAC)

for active voltage control, which can couple with mainstream multi-agent actor-critic algorithms. Specifically, we propose a transformer-based policy network and a transformer-based critic network to obtain discriminative representations. For the policy network, we divide the whole observation into node-based entities and project these node-based entities to high-dimensional semantic space via a transformer encoder. Inspired by [22, 29], an adjacency matrix representing the connectivity of nodes in the zone is treated as the mask in the self-attention mechanism, introducing topological information to enhance representations. Then, we propose an embedding aggregation module to aggregate node-based information in the zone into the embedding from the agent’s (control unit) point of view. To address the instability of transformer architecture in MARL algorithms[14, 22], we develop a novel self-supervised auxiliary task in the training process of policy networks. For the critic network, we exploit the vanilla transformer layer to approximate the global Q-value function. We introduce the self-attention mechanism to model the correlations between agents from the scale of the entire grid. Experiments on the MAPDN environment[26] demonstrate the effectiveness of our approach.

In summary, our main contributions are three-fold as follows:

- We propose a novel transformer-based multi-agent actor-critic framework for active voltage control task and improve the performances of the existing multi-agent actor-critic algorithms from the perspective of voltage regulation and energy loss.
- We adopt a self-supervised auxiliary task to stabilize the training process of MARL algorithms with transformer, improving the sample efficiency and facilitating the representation learning.
- We introduce the attention mechanism into the voltage control in power network tasks, assisting a control unit in elaborating cooperative control strategies with other control units. It is more explainable and facilitates the MARL-based methods deploying to the realistic power system.

The rest of this paper is organized as follows. We first give a literature review on the related work in Section 2. Then the background of active voltage control and the formulation of Markov Games are elaborated in Section 3. The methodology is described in detail in Section 4. Simulation results and discussions are presented in Section 5. Finally, we conclude our work in Section 6.

## 2 RELATED WORKS

### 2.1 Multi-Agent Reinforcement Learning for Active Voltage Control

Advances in machine learning lead to widespread applications of multi-agent reinforcement learning techniques to tackle active voltage control problem. In [4], authors take advantage of spectral clustering algorithms to partition the large distribution power network into several zones and formulate the control between each zones as Markov Game solved by MATD3. [3] introduced an attention mechanism in the critic network to enhance scalability of algorithms. In contrast to [3], we not only introduce attention mechanisms to model the relationship of agents in the critic network, but also propose transformer-based architecture adapting to the grid topology in the policy network. [27] developed a approach with a manually designed voltage inner loop for the autonomous voltage control of transmission network based on MADDPG. [5]

leverages the sparse pseudo-Gaussian process to build a surrogate model using few-shot recorded data and control actions based on multi-agent soft actor critic algorithms. Above prior works divided the whole grid into several zones, with each zone controlled by a single agent. However, MAPDN[26] modeled the active voltage control problem as a Dec-POMDP[21] and each PV inverter is controlled by an agent. In this paper, we also follow the settings of Dec-POMDP proposed in MAPDN, which enables the presence of multiple agents with similar observations in a zone.

## 2.2 Attention Mechanisms and Transformer in Multi-Agent Reinforcement Learning

Transformers[25] have been applied successfully to solve a wide variety of tasks in natural language processing[8] and computer vision[6, 9]. However, transformers have not yet been fully explored in multi-agent reinforcement learning, mostly due to differing nature of problem, such as high variance in training. UPDet[14] proposed the transformer-based individual value function and policy decoupling strategy based on value-based MARL algorithms to achieve improvements on multi-agent games in the StarCraft II environment with discrete action space. Hierarchical RNNs-Based Transformers MADDPG (HRTMADDPG)[28] combined transformers with RNNs, capturing the causal relationship between sub-time sequences. [17] proposed a self-attention-based multi-agent continuous control algorithms to solve the problem of uneven learning degree and improved learning efficiency when faced with more agents. The above works showed promising performance on game-like environments with a few agents. However, it is not yet clear whether multi-agent algorithms with transformer can still achieve competitive performance if applied to the power system with large-scale agents, i.e. there are 38 agents on 322-bus network in MAPDN environment[26].

## 3 PROBLEM FORMULATION

### 3.1 Active Voltage Control on Power Distribution Networks

In this paper, a power distribution network installed with roof-top photovoltaics (PVs) is modeled as a tree graph structure  $\mathcal{G} = (V, E)$  shown in Figure 1, where  $V = \{0, 1, \dots, N\}$  and  $E = \{1, 2, \dots, N\}$  denote the set of nodes (buses) and edges (branches), respectively[11]. For bus  $i \in V$ , let  $v_i$  and  $\theta_i$  be the magnitude and phase angle of complex voltage and  $s_i = p_i + j q_i$  denotes the complex power injection. There are complex and non-linear relationships between these physical quantities that satisfy the power system dynamics rules[26]. In particular, node 0 is connected to the main grid, which serves to balance the active and reactive power in the distribution network. Nodes in the distribution network are divided into several zones based on their shortest path from the terminal to the main branch[26]. Also, loads (e.g. residential and industrial clients) and PVs are connected to some of nodes. Each PV is equipped with an PV inverter that generates reactive power to control the voltage around the standard value denoted as  $V_{ref}$ . For safe operation of the distribution network, 5% voltage fluctuation is usually allowed. Let  $v_0 = 1.0$  per unit (p.u.), the voltage amplitude of each bus must satisfy the following inequality condition:

$0.95p.u. \leq v_i \leq 1.05p.u., \forall i \in V \setminus \{0\}$ . In the middle of day, the solar energy is converted into electrical energy and injected into the distribution network via PVs, which would increase  $v_i$  out of the safe range. In contrast,  $v_i$  may drop below the 0.95p.u. due to the heavy load at night. In this paper, we consider each PV inverter installed in a PV as the control unit.

### 3.2 Formulation of Markov Games in Active Voltage Control

The collaborative control process of PV inverters can be modeled as a Dec-POMDP[21] for  $N$  agents. A Dec-POMDP is usually defined by a tuple  $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \mathcal{T}, r, \{O_i\}_{i \in \mathcal{N}}, \Omega, \gamma)$ , where  $\mathcal{N} = \{1, \dots, n\}$  denotes the set of  $n$  agents,  $\mathcal{S}$  denotes the state space observed by all agents,  $\mathcal{A}_i$  denotes the action space of agent  $i$ . Let  $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}_i$ , then  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  denotes the transition probability from any state  $s \in \mathcal{S}$  to any state  $s' \in \mathcal{S}$  after taking a joint action  $a \in \mathcal{A}$ ;  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is a global reward function that determines the immediate reward received by whole agents for a transition from  $(s, a)$  to  $s'$ ;  $O = \times_{i \in \mathcal{N}} O_i$  denotes the joint observation set, where  $O_i$  is each agent's observation;  $\Omega : \mathcal{S} \times \mathcal{A} \times O \rightarrow [0, 1]$  denotes the perturbation of the observers for agents' joint observations over the states after decisions;  $\gamma \in [0, 1)$  is the discount factor. We can formulate the policy of the  $i$ -th agent's policy as  $\pi^i$ , and the objective of Dec-POMDP is finding an optimal joint policy  $\pi = \times_{i \in \mathcal{N}} \pi^i$  to maximize expected long-term reward  $\mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r_t]$ . Considering the active voltage control problem, we describe specific elements in the Dec-POMDP in detail as follows, simulated to [26]:

- **Agent.** As shown in Figure 1, each PV is an agent that injects the reactive power generated by its PV inverter into the distribution network so as to maintain the voltage of all buses within the safe range.
- **Observation.** Let  $o_j = (p_j^L, q_j^L, v_j, \theta_j, flag_j^{PV})$  represents the node-based feature of node  $j$ .  $p_j^L \in (0, \infty)$  and  $q_j^L \in (0, \infty)$  are active and reactive power of the load connected to node  $j$  respectively;  $v_j \in (0, \infty)$  and  $\theta_j \in [-\pi, \pi]$  denote the voltage magnitude and voltage phase of node  $j$  respectively;  $flag_j^{PV} \in \{0, 1\}$  indicates whether the node  $j$  is installed with a PV. Moreover, additional physical quantities  $(p^{PV}, q^{PV})$  are appended to the node-based feature for those nodes installed with PV.  $p^{PV} \in (0, \infty)$  denotes the active power generated by PV  $i$  and  $q^{PV} \in (-\infty, \infty)$  is the reactive power generated by the PV inverter. Nodes in the grid are partitioned into several zones and the observation of agent  $i$  (denoted as  $O_i$ ) is obtained by concatenating node-based features in the zone in which agent  $i$  is located. It is worth noting that agents in the same zones have similar observations.
- **Action.** Each agent  $i \in \mathcal{N}$  has a continuous action set  $\mathcal{A}_i = \{a_i : -c \leq a_i \leq c, c > 0\}$  that denotes the ratio of maximum reactive power it can generate. And the joint action set is defined as  $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}_i$ .
- **Reward Function.** The reward function is defined as follows:

$$r = -\frac{1}{|V|} \sum_{i \in V} l_v(v_i) - \alpha \cdot l_q(q^{PV}), \quad (1)$$

where  $l_v(\cdot)$  is a voltage barrier function and  $l_q = \frac{1}{|N|} \|\mathbf{q}^P V\|_1$  is the reactive power generation loss. The objective is to learn a optimal strategy to control the voltage within the safety range around  $V_{ref}$  (i.e.  $[0.95v_{ref}, 1.05v_{ref}]$ ) while minimizing the power loss of the whole distribution network. The hyper-parameter  $\alpha \in (0, 1)$  is set in advance by simulation environment, which plays the role of balancing the two losses. In practice, we use the reactive power generation loss instead of power loss of the whole distribution network, because obtaining overall power loss of the grid in real time is difficult. The voltage barrier function  $l_v(\cdot)$  penalizes the voltage rise deviation and the voltage drop deviation. As with [26], three voltage barrier functions are considered to form reward functions: L1-shape, L2-shape and bowl-shape (see Appendix A.3).

Additionally, the topology of the grid will be exploited as a priori knowledge in our framework. Formally, the connectivity between nodes in a zone is represented by a adjacency matrix  $D^i$ . Furthermore, we divide the raw observation of agent  $i$  into node-based features based on the prior knowledge:

$$O_i = \{o_{i,1}, o_{i,2}, \dots, o_{i,m_i}\}. \quad (2)$$

Here,  $i \in \{1, 2, \dots, n\}$  is the index of agents,  $m_i$  denotes the number of nodes in the zone. Let  $p_i \in \{1, 2, \dots, m_i\}$  denotes the index of node installed with PV  $i$ , which makes agent  $i$  aware of its own location in the zone.

## 4 METHOD

In order to apply MARL algorithms to the active voltage control problem while considering the characteristics of grid, we present a novel transformer-based multi-agent actor critic (T-MAAC) framework. Specifically, we propose a policy network and a critic network based on transformer as well as auxiliary-task based training process, which is compatible with mainstream multi-agent actor-critic algorithms such as MADDPG[19] and MATD3[1]. In this section, we describe the structure of the policy network and the critic network in Section 4.1 and Section 4.2, respectively. The auxiliary task to stabilize the training process is discussed in Section 4.3.

### 4.1 Transformer-based Policy Network

To extract more relevant representations for the active voltage control task, we develop a transformer-based policy network as shown in Figure 2 to handle various types of observations. We present a mathematical formulation of our transformer-based model in this section.

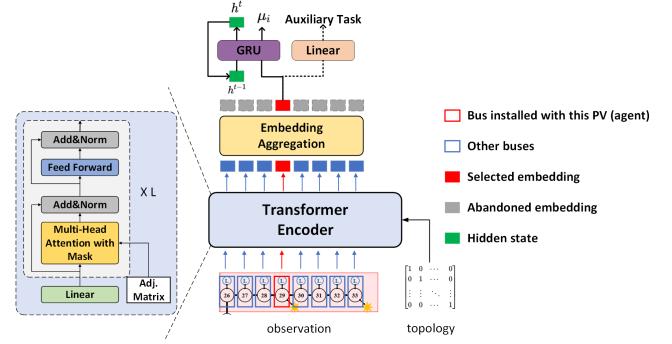
**4.1.1 Projection layer.** First of all, we transform the raw observation  $O$  into node-based embeddings via a projection layer.

If the observation of agent  $i$  at time step  $t$  (denoted as  $O_i^t$ ) is made up of  $m$  node-based features, then all of them are embedded via a projection layer  $P$  as follows:

$$\text{input}_i^t = \{P(o_{i,1}^t), P(o_{i,2}^t), \dots, P(o_{i,m_i}^t)\}. \quad (3)$$

In Eq. (3),  $i \in \{1, \dots, n\}$  is the index of the agent;  $j \in \{1, \dots, m_i\}$  is the index of nodes;  $m$  denotes the number of nodes in the zone.

In the vanilla transformer[25], Vaswani adds "positional encodings" to the input embeddings to inject position information of the tokens in a sequence. However, a distribution network has a radial



**Figure 2: The architecture of the policy network based on transformer.** The transformer encoder and embedding aggregation module are used to obtain the embedding of raw observation. Then, the embedding is mapped to action via the GRU head. Additionally, the auxiliary-task head predicts the voltage out of control ratio (VR) of the raw observation, which stabilizes the training process via extra auxiliary loss. Details can be found in Section 4.1.

topology instead of sequential structure. So instead of positional encodings, we inject position information via an adjacency matrix  $D^i$  used in the attention mechanism that will be elaborated in the next section.

**4.1.2 Transformer Encoder and Embedding Aggregation Module.** Next, The transformer encoder and embedding aggregation module are designed to extract more robust representation from the  $\text{input}_i^t$  above as shown in Figure 2.

The vanilla self-attention mechanism proposed in [25] is computed as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V}. \quad (4)$$

Eq. (4) can be extended to self-attention with mask mechanism:

$$\text{MaskAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \text{mask}) = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \cdot \text{mask} \right) \mathbf{V}. \quad (5)$$

Three matrices  $\mathbf{K}$ ,  $\mathbf{Q}$ ,  $\mathbf{V}$  represent a set of keys, queries and values respectively;  $d_k$  is a scaling factor equal to dimension of queries and keys;  $\text{mask}$  is a binary matrix with the same shape as  $\mathbf{Q}\mathbf{K}^T$  and  $\text{mask}_{x,y} \in \{0, 1\}$  indicates whether perform an attention operation between position  $x$  and position  $y$ .

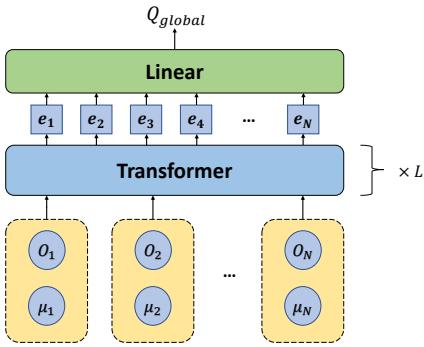
Our transformer encoder utilizes the self-attention with mask mechanism to establish the correlation between nodes in the zone. We formulate our transformer encoder as follows:

$$E^0 = \text{input}_i^t, \quad (6)$$

$$Q^{(l)}, K^{(l)}, V^{(l)} = W_Q^{(l)} E^{(l-1)}, W_K^{(l)} E^{(l-1)}, W_V^{(l)} E^{(l-1)}, \quad (7)$$

$$\bar{Y}^{(l)} = \text{MaskAttention} \left( Q^{(l)}, K^{(l)}, V^{(l)}, D^i \right), \quad (8)$$

$$E^{(l)} = \text{LayerNorm} \left( E^{(l-1)} + \text{Linear} \left( \bar{Y}^{(l)} \right) \right), \quad (9)$$



**Figure 3: The architecture of the global critic network. Our framework replace the widely used MLP-based critic network with the transformer-based critic network.**

where  $W_Q^{(l)}, W_K^{(l)}, W_V^{(l)}$  represent the learnable parameters to compute  $Q, K, V$  and  $l \in \{1, 2, \dots, L\}$  is the index of transformer layers. The adjacency matrix  $D^i$  indicates how nodes are connected. For example, if node  $x$  and node  $y$  are adjacent,  $D_{xy}^i$  equals 1; if they are not,  $D_{xy}^i$  equals 0.

Then, the embedding aggregation module aggregates  $E^{(l)}$  into global information  $\bar{E}_i$  from the view of agent  $i$ . In practice, we select an additional transformer layer without mask as our embedding aggregation module:

$$Q^{EA}, K^{EA}, V^{EA} = W_Q^{EA} E^{(l)}, W_K^{EA} E^{(l)}, W_V^{EA} E^{(l)}, \quad (10)$$

$$\bar{Y}^{EA} = \text{Attention}(Q^{EA}, K^{EA}, V^{EA}), \quad (11)$$

$$E^{EA} = \text{LayerNorm}\left(E^{(l)} + \text{Linear}(\bar{Y}^{EA})\right), \quad (12)$$

$$\bar{E}_i = \text{SelectEmbedding}(E^{EA}, p_i). \quad (13)$$

SelectEmbedding is an operation to select the embedding whose index is  $p_i$  (The index of node installed with PV  $i$ ). Intuitively,  $\bar{E}_i$  denotes the representation extracted from the agent  $i$  perspective in a higher semantic space.

**4.1.3 GRU head and Auxiliary-task head.** In the last part of the policy network, a GRU[7] layer is applied to project the representation  $\bar{E}_i$  to the control action  $\mu_i$ .  $h^t$  in Figure 2 denotes the temporal hidden state at the time step  $t$ .

Meanwhile, we introduce the auxiliary-task head to predict the voltage out of control ratio (VR) in  $O_i$ . VR indicates the ratio of voltage outside the safety range (i.e. 0.95–1.05 p.u.) in a zone, which is a critical metric that corresponds to the optimization objectives for active voltage control task. Thus, the auxiliary-task head helps the upstream transformer encoder module implicitly recover this essential information from raw observation. Also, extra auxiliary loss helps stabilize the learning process, which is introduced in Section 4.3 in detail.

## 4.2 Transformer-based Critic Network

For the active voltage control task, the correlations between agents are related to their position in the grid. In a radial distribution

network, the voltage of each node is influenced by all other nodes, but the impact decreases as the distance increases. Therefore, agents need to be aware of the relationship between nodes in the grid to make cooperative control decisions. For example, if two agents are topologically close to each other, they must consider more carefully when making decisions. Additionally, the voltage of most nodes can be controlled within the safety range (i.e. from 0.95 p.u. to 1.05 p.u.), while only a few parts have the risk of voltage exceeding the safety value. Thus, agents may pay more attention to those zones in danger to control the voltage of all nodes in the distribution network.

The conventional centralized critic network constructed with pure MLPs[26] suffers from credit assignment issues, especially in a large-scale cooperative environment. The large number of agents and their complex relationships complicate policy learning[18]. To make the learning process more robust, we model the correlation between agents via the self-attention mechanism and design a transformer-based critic network. The architecture of critic network is shown in Figure 3.

We describe the calculation of the global Q-function parameterized by the transformer-based critic network. First of all, let  $O_i$  denotes the raw observation of agent  $i$  and  $\mu_i$  denotes the corresponding action it performed. Then,  $N$  tuples  $(O_i, \mu_i)$  are transformed into  $N$  embeddings through several vanilla transformer layers[25]:

$$\{e_1, \dots, e_N\} = \text{Transformer}(\{(O_1, \mu_1), \dots, (O_N, \mu_N)\}). \quad (14)$$

Next, we project the embeddings to the output space of the centralized action-value function  $Q_{global}^\pi$  through a linear function:

$$Q_{global}^\pi(O_1, \mu_1, O_2, \mu_2, \dots, O_N, \mu_N) = \text{Linear}(e_1, e_2, \dots, e_N). \quad (15)$$

In addition, our proposed transformer-based critic network reduces the number of parameters that benefits from compact semantic representations calculated by transformer architecture.

## 4.3 Auxiliary-task Training Process

In this section, we select MADDPG[19] as the base algorithm to describe the entire auxiliary-task training process of T-MAAC. Suppose there are a total of  $N$  agents with continuous policies  $\mu_i(\cdot; \theta_i)$  parameterized by  $\theta = \{\theta_1, \dots, \theta_N\}$ . Let  $s = (o_1, \dots, o_N)$  denotes the observations of each agents, then we formulate the gradient of the expected return for agent  $i$ ,  $J(\mu_i) = \mathbb{E}[R]$  as:

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}_{s, a \sim \mathcal{D}} \left[ \nabla_{\theta_i} \mu_i(a_i | o_i) \nabla_{a_i} Q_{global}^\mu(s, a_1, \dots, a_N) \Big|_{a_i=\mu_i(o_i)} \right]. \quad (16)$$

Here,  $\mathcal{D}$  is the experience replay buffer recording transitions of all agents. And the centralized global action-value functions  $Q_{global}^\mu$  is updated as:

$$y = r + \gamma Q_{global}^{\mu'}(s', a'_1, \dots, a'_N) \Big|_{a'_j=\mu'_j(o_j)}, \quad (17)$$

$$\mathcal{L}(\theta_i) = \mathbb{E}_{s, a, r, s'} \left[ \left( Q_{global}^\mu(s, a_1, \dots, a_N) - y \right)^2 \right]. \quad (18)$$

In addition, we adopt an additional self-supervised loss to stabilize the learning process.  $v_i(\cdot; \theta_i)$  is the output of auxiliary-task head in the policy network, which predicts the voltage out of control

ratio ( $VR$ ) in the zone. And it's ground truth  $label_i$  is calculated from the raw observation  $o_i$ . The auxiliary loss is as follows:

$$label_i = \sum_{j=1}^{m_i} [\mathbb{I}(Voltage_j < 0.95) + \mathbb{I}(Voltage_j > 1.05)], \quad (19)$$

$$\mathcal{L}_{aux}(\theta_i) = \mathbb{E}_{s,a,r,s'} [(\mathbf{v}_i(o_i; \theta_i) - label_i)^2], \quad (20)$$

where  $\mathbb{I}$  is the indicator function and  $m_i$  is the number of nodes in the  $o_i$ . In order to improve the sample efficiency and scalability in MARL algorithms, the parameters of policy networks are shared among agents in T-MAAC. Details of transformer-based MADDPG (T-MADDPG) are given in Algorithm 1.

---

**Algorithm 1:** Transformer-based MADDPG (T-MADDPG)

---

```

1 for episode  $\leftarrow 1$  to  $M$  do
2   Initialize a random process  $\mathcal{N}$  for action exploration and
      Replay Buffer  $\mathcal{D}$ ;
3   for  $t \leftarrow 1$  to max-episode-length do
4     for each agent  $i$ , select a action  $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_i$ ;
5     Execute actions  $a = (a_1, \dots, a_N)$  from state  $s$ ;
6     Get reward  $r$  by going to new state  $s'$ ;
7     Store  $(s, a, r, s')$  in replay buffer  $\mathcal{D}$  and  $s \leftarrow s'$ ;
8     for agent  $i \leftarrow 1$  to  $N$  do
9       Sample a random mini-batch of  $S$  samples
           $(s^j, a^j, r^j, s'^j)$  from  $\mathcal{D}$  ;
10      Update critic by minimizing the loss in Eq. (18) ;
11      Update actor by minimizing the auxiliary loss in
          Eq. (20) ;
12      Update actor by gradient ascent in Eq. (16) ;
13    end
14    Update target network parameters for each agent  $i$ :
15     $\theta'_i \leftarrow \tau\theta_i + (1 - \tau)\theta'_i$ ;
16  end

```

---

## 5 EXPERIMENTS

In this section, we conduct a series of experiments based on the MAPDN environment[26] to evaluate the performance of T-MAAC. We first introduce the experiment setups and implementation details of the algorithms. Then, we compare the evaluation results of our algorithm with the baseline methods and the ablated variants. Furthermore, we show a case study that visualizes the attention weights in the self-attention mechanism to analyze which nodes in the distribution network agents should focus on (see Appendix B.1).

### 5.1 Experiment Setups

The MAPDN[26] is an environment of distributed/decentralized active voltage control on power distribution networks, which supports numerical studies for the 33-bus, 141-bus, and 322-bus network. We conduct experiments in the 141-bus network scenario with 22 agents and the 322-bus network scenario with 38 agents.

In order to evaluate different algorithms fairly, we randomly select some test scenarios from different seasons to construct a test dataset and a validation dataset(details in Appendix A.1). Each experiment is run with 5 random seeds and the test results during training are given by the median and the 25%-75% quartile shading. And each experiment is evaluated in the validation dataset every 20 episodes during training. After the training phase, we evaluate the learned strategy on the whole test dataset.

Following the proposal of [26], we conduct main experiments with three different voltage barrier functions (see Appendix A.3) on the 141-bus and the 322-bus networks. In experiments, we use two metrics to evaluate the performance of algorithms. *Controllable rate (CR)*: It calculates the ratio of all buses' voltage being under control within safety range. *Q loss (QL)*: It calculates the mean reactive power generations by agents per time step, which is the same as  $l_q(\cdot)$  defined in Eq.(1). QL is an alternative metric to power loss because the power loss of the whole grid is hard to obtain in the actual distribution network. CR is the most critical metric for the active voltage control task, and a low CR indicates that the entire grid is currently perilous.

### 5.2 Baseline Methods and Implementation Details

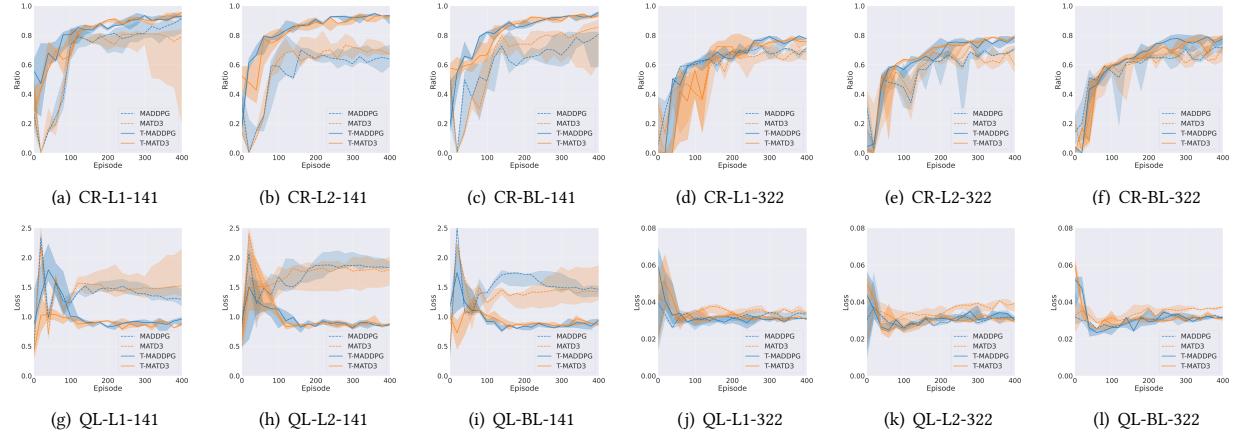
According to [26], MADDPG[19] and MATD3[1] achieve excellent performance in the MAPDN environment compared to other state-of-the-art MARL algorithms. Thus, we separately couple our proposed T-MAAC with MADDPG and MATD3 to evaluate the performance of our framework:

- **MADDPG and MATD3.** In our experiments, the MLP-based policy network in MADDPG and MATD3 consists of one hidden layer and a GRU layer. And the MLP-based critic network is constructed with one hidden layer.
- **T-MADDPG and T-MATD3.** Compared to the baseline algorithms above, T-MADDPG and T-MATD3 replace the MLP-based network architecture with the transformer-based network architecture proposed in Section 4.1 and 4.2. Moreover, additional auxiliary loss is introduced during training (see Algorithm 1). In the policy network, the transformer encoder is composed of a stack of 2 transformer layers, and the embedding aggregation module is an extra transformer layer. The architecture of GRU head remains the same as the baseline algorithms above.

Following [13], all algorithms are trained with the normalized reward and the action bound enforcement trick. We perform gradient clipping with L1 norm and the clip bound is set to 1. Moreover, the parameters in policy networks are shared among agents, and the agent ID is concatenated with observation to distinguish different agents. The hyper-parameters of algorithms are shown in Table 2 (see Appendix A.4).

### 5.3 Result

The median CR and QL of algorithms during training for MADDPG, MATD3, T-MADDPG and T-MATD3 are shown in Figure 4. As the figure shows, our proposed T-MAAC framework consistently improves the performance of baseline methods under three types of rewards and two different scale grid scenarios. Owing to the learned



**Figure 4: Median CR and QL of algorithms with different voltage barrier functions. "T-" indicates the combination of T-MAAC and the baseline algorithm. The sub-caption indicates metric-Barrier-scenario and BL is the abbreviation of bowl.**

**Table 1: The mean test results in the test dataset. CR denotes the control rate; QL denotes the Q loss; PL denotes the power loss.**

Method	Spring			Summer			Fall			Winter		
	CR (%)	QL ( $\frac{\text{MW}}{\text{MVAR}}$ )	PL (MW)	CR (%)	QL ( $\frac{\text{MW}}{\text{MVAR}}$ )	PL (MW)	CR (%)	QL ( $\frac{\text{MW}}{\text{MVAR}}$ )	PL (MW)	CR (%)	QL ( $\frac{\text{MW}}{\text{MVAR}}$ )	PL (MW)
322-MADDPG	79.2	0.031	0.036	75.0	0.031	0.040	92.5	0.031	0.029	97.8	0.031	0.027
322-T-MADDPG	88.6	0.027	0.037	86.3	0.027	0.044	96.7	0.025	0.026	98.0	0.024	0.021
322-MATD3	59.4	0.033	0.037	54.1	0.033	0.040	77.9	0.034	0.033	89.4	0.035	0.033
322-T-MATD3	90.6	0.028	0.039	89.5	0.029	0.046	97.9	0.028	0.028	99.3	0.027	0.024
141-MADDPG	75.8	1.88	0.78	72.9	1.85	0.91	75.4	1.95	0.53	71.1	1.99	0.45
141-T-MADDPG	97.8	0.77	0.93	97.8	0.78	1.10	100	0.79	0.61	100	0.85	0.50
141-MATD3	78.7	1.83	0.93	73.4	1.77	1.08	90.1	1.95	0.67	89.9	2.03	0.59
141-T-MATD3	97.4	0.77	0.89	97.4	0.79	1.06	99.9	0.79	0.57	100	0.85	0.47

better representations relevant to the active voltage control task, T-MADDPG and T-MATD3 improve the controllable rate (CR) in the grid while reducing the reactive power generation (QL). T-MAAC significantly performs well on the 141-bus network, verifying the importance of better representations. As for the 322-bus network, a large-scale scenario with 38 agents, T-MADDG increases the CR while maintaining the low QL as same as MADDPG, and T-MATD3 achieves a higher CR with a lower QL. By comparing the performance with three different reward functions, T-MAAC also stabilizes the training process and alleviates the phenomenon mentioned in [26] that algorithms is sensitive to reward functions. The better representations improve sample efficiency and resolve the issue that different reward functions lead agents to different local optimal policies.

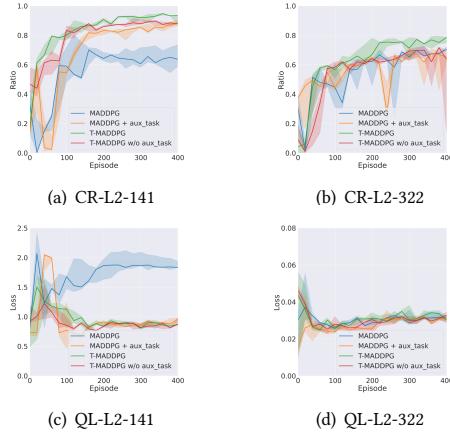
We also evaluate all algorithms trained by L2-shape voltage barrier function in the test dataset. We show the performances in Table 1 (including actual power loss *PL* in the whole distribution network). The metric *CR* shows that it is more challenging to control the voltage in spring and summer than in fall and winter due to excessive active power injection produced by PVs in the former. Our

proposed T-MAAC overcomes such difficulties and achieves better performances in all scenarios, especially in spring and summer. Meanwhile, penalized by *Q loss*, T-MAAC learns a strategy with less reactive power generation than baseline methods. It is worth noting that *QL* is a proxy for power loss during training, thus, the learned strategy with less *QL* may still lead to more *PL*. Such a matter can be alleviated by more related and easy-to-obtain surrogate metric[5].

## 5.4 Ablation Study

In this section, we conduct a series of experiments to examine further which particular components of T-MAAC are essential for the performance. We fix the baseline algorithm to be MADDPG trained by L2-shape voltage barrier function. The performances of the variants of T-MADDPG are evaluated in the following experimental settings.

**5.4.1 Auxiliary Task.** We add an additional auxiliary task during training to stabilize the training process (see Section 4.3). The following ablated variants are designed to verify the effectiveness of the auxiliary task:



**Figure 5: Performance comparison with or w/o auxiliary task. The sub-caption indicates metric-Barrier-scenario.**

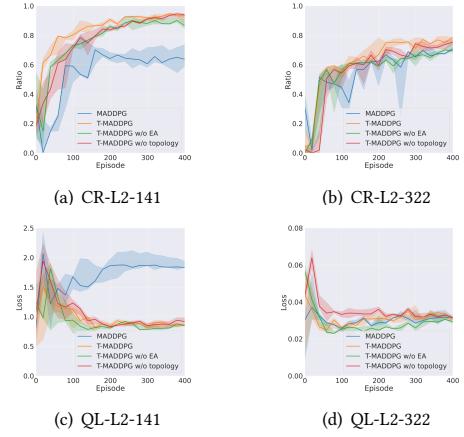
- **T-MADDPG w/o aux task:** We remove the auxiliary task and optimize model by Eq. (16) and Eq. (18).
- **MADDPG with aux task:** We also add the auxiliary task to the MLP-based MADDPG to figure out whether the auxiliary task can improve performance with different network architectures. The result is shown in Figure 5. The performance of T-MADDPG is better than T-MADDPG without auxiliary task during almost the entire training process especially on the 322-bus network that is a more challenging large-scale scenario. T-MADDPG without auxiliary task performs little differently from MADDPG, which indicates training with auxiliary task is essential for our transformer-based network architectures.

In addition, training with auxiliary task also improves the performance of MADDPG on the 141-bus network, which means that the auxiliary loss during training helps convergence to a better policy. However, training with auxiliary task doesn't work on the 322-bus network. This may be due to the fact that the simple policy/critic network constructed with MLPs is no longer able to capture the nature in the grid, especially in the large-scale grid.

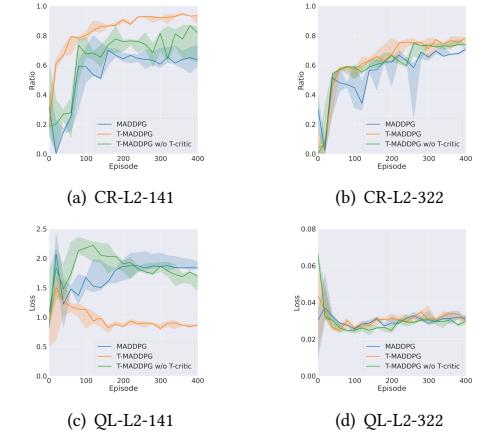
**5.4.2 Ablation Study on the Transformer-based Policy Network.** In T-MAAC, we design a novel policy network based on transformer to achieve better performance on active voltage control task (see Section 4.1). We introduce two ablated variants to examine modules in the transformer-based policy network as follows:

- **T-MADDPG w/o EA:** We remove the embedding aggregation module, and the final representation of  $O$  is obtained by average outputs of the transformer encoder.
- **T-MADDPG w/o topology:** We remove the adjacency matrix and the mask in the self-attention mechanism. Attention operations in Eq. (4) are implemented between all nodes regardless of whether they are connected in the zone or not.

The embedding aggregation module integrate node-based information to global information from the point of the agent, and the result in Figure 6 shows that it further improves performance on both 141-bus network and 322-bus network. Further more, by



**Figure 6: Performance comparison on variants of our policy network. The sub-caption indicates metric-Barrier-scenario.**



**Figure 7: Performance comparison on variants of our critic network. The sub-caption indicates metric-Barrier-scenario.**

comparing the performance of T-MADDPG w/o EA and MADDPG in the 322-bus network, it can be seen that vanilla transformer architecture without EA can't handle various observation space. This result verifies the opinions discussed in Section 4.1 that aggregating information from the perspective of decision maker allows transformer encoder to obtain better representations suitable for this task.

If we don't inject the position information into the transformer, all nodes are treated equally in the early stages of training. Thus, inspired by [22, 29], we select the adjacency matrix as the mask in the self-attention mechanism to assist agents in capturing the nature of the grid. As shown in Figure 6, the utilization of topology improves sample efficiency in the baseline algorithm.

**5.4.3 Transformer-based Critic.** In this section, we conduct ablation experiments to figure out the effect of introducing transformers into the global critic network.

- **T-MADDPG w/o T-critic:** We replace the transformer-based critic network with the widely used MLP-based critic network as same as MADDPG.

We show the result in Figure 7. Compared to T-MADDPG w/o transformer-based critic, T-MADDPG achieves better performance on both scenarios. Moreover, the transformer-based critic also stabilizes the training process in 141-bus network as shown in Figure 7(b).

## 6 CONCLUSIONS

In this paper, we propose T-MAAC, a transformer-based multi-agent actor-critic framework, for voltage stabilization in power distribution networks. Our framework consists of a policy network and a global critic network. The policy network based on transformer captures the characteristics of grid, obtaining better representations for power network task. In the global critic network, we introduce the self-attention mechanism to model the correlation between agents and achieving better performance. Additionally, we adopt the auxiliary task, predicting the voltage out of control ratio in a zone for active voltage control task, to stabilize the training process and improve the embedding learning. We conduct extensive evaluations as well as ablation studies in the real-world scale grid scenarios provided by MAPDN. The experimental results demonstrate that T-MAAC significantly improves the performance of existing MARL algorithms for voltage stabilization.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Contract 61836011 and in part by the Youth Innovation Promotion Association CAS under Grant 2018497. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

## REFERENCES

- [1] Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. 2019. Reducing overestimation bias in multi-agent domains using double centralized critics. *arXiv preprint arXiv:1910.01465* (2019).
- [2] Mesut E Baran and Felix F Wu. 1989. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Power Engineering Review* 9, 4 (1989), 101–102.
- [3] Di Cao, Weihao Hu, Junbo Zhao, Qi Huang, Zhe Chen, and Frede Blaabjerg. 2020. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters. *IEEE Transactions on Power Systems* 35, 5 (2020), 4120–4123.
- [4] Di Cao, Junbo Zhao, Weihao Hu, Fei Ding, Qi Huang, and Zhe Chen. 2020. Distributed voltage regulation of active distribution system based on enhanced multi-agent deep reinforcement learning. *arXiv preprint arXiv:2006.00546* (2020).
- [5] Di Cao, Junbo Zhao, Weihao Hu, Fei Ding, Qi Huang, Zhe Chen, and Frede Blaabjerg. 2021. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs. *IEEE Transactions on Smart Grid* 12, 5 (2021), 4137–4150.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision*. Springer, 213–229.
- [7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [10] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [11] Lingwen Gan, Na Li, Ufuk Topcu, and Steven H Low. 2013. Optimal power flow in tree networks. In *Proceedings of the IEEE Conference on Decision and Control*. IEEE, 2313–2318.
- [12] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multi-agent control using deep reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*. Springer, 66–83.
- [13] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905* (2018).
- [14] Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. 2021. UPDeT: Universal Multi-agent RL via Policy Decoupling with Transformers. In *International Conference on Learning Representations*.
- [15] HM Khodr, FG Olsina, PM De Oliveira-De Jesus, and JM Yusta. 2008. Maximum savings approach for location and sizing of capacitors in distribution systems. *Electric Power Systems Research* 78, 7 (2008), 1192–1203.
- [16] Haotian Liu and Wenchuan Wu. 2021. Online multi-agent reinforcement learning for decentralized inverter-based volt-var control. *IEEE Transactions on Smart Grid* 12, 4 (2021), 2980–2990.
- [17] Kai Liu, Yuyang Zhao, Gang Wang, and Bei Peng. 2022. Self-attention-based multi-agent continuous control method in cooperative environments. *Information Sciences* 585 (2022), 454–470.
- [18] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. 2020. Multi-agent game abstraction via graph attention neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 7211–7218.
- [19] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275* (2017).
- [20] Steffen Meinecke, Džana Sarajlić, Simon Ruben Drauz, Annika Klettke, Lars-Peter Lauven, Christian Rehntanz, Albert Moser, and Martin Braun. 2020. Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energy* 13, 12 (2020), 3290.
- [21] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.
- [22] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Séb Noury, et al. 2020. Stabilizing transformers for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 7487–7498.
- [23] Leon Thurner, Alexander Scheidler, Florian Schäfer, Jan-Hendrik Menke, Julian Dollichon, Friederike Meier, Steffen Meinecke, and Martin Braun. 2018. pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems* 33, 6 (2018), 6510–6521.
- [24] Tijmen Tielemans, Geoffrey Hinton, et al. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* 4, 2 (2012), 26–31.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*. 5998–6008.
- [26] Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C Green. 2021. Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems* 34 (2021), 3271–3284.
- [27] Shengyi Wang, Jiajun Duan, Di Shi, Chunlei Xu, Haifeng Li, Ruisheng Diao, and Zhiwei Wang. 2020. A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning. *IEEE Transactions on Power Systems* 35, 6 (2020), 4644–4654.
- [28] Xiaolong Wei, Xianglin Huang, LiFang Yang, Gang Cao, Zhulin Tao, Bing Wang, and Jing An. 2021. Hierarchical RNNs-Based transformers MADDPG for mixed cooperative-competitive environments. *Journal of Intelligent & Fuzzy Systems* Preprint (2021), 1–12.
- [29] Deunsol Yoon, Sungsoon Hong, Byung-Jun Lee, and Kee-Eung Kim. 2020. Winning the L2RPN challenge: Power grid management via semi-markov afterstate actor-critic. In *International Conference on Learning Representations*.

- [30] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2021. The surprising effectiveness of mappo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955* (2021).

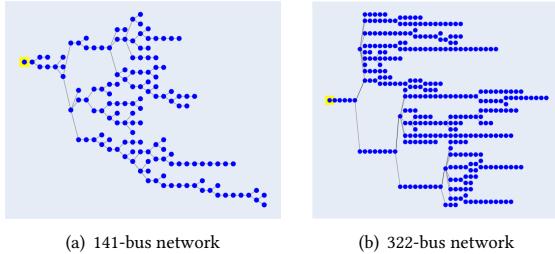
## A EXPERIMENTAL SETTINGS

### A.1 MAPDN Environment and Datasets

In MAPDN, the 33-bus and 141-bus networks are modified from IEEE 33-bus[2] and IEEE 141-bus[15], respectively, while the 322-bus network is constructed by topology from SimBench[20]. The data supporting simulation (i.e., load profile, load and PV data, active and reactive power consumption) are collected from the real world and then interpolated with the 3-min resolution consistent with the grid’s real-time control period. To guarantee the safety of distribution network, the environment manually sets the range of actions with  $[-0.6, 0.6]$  for the 141-bus case and  $[-0.8, 0.8]$  for the 322-bus case suggested by MAPDN[26]. As for the reward function,  $\alpha$  in Eq.(1) is set to 0.1 in the 322-bus case and 0.01 in the 141-bus case to tune the trade-off between voltage control performance and reactive power generation loss.

It is worth noting that the difficulty of voltage control problem varies during different months of a year. For example, during the midday summer, excessive active power from intense sunlight is injected into the grid, creating a more significant challenge for the voltage control task than in winter. Thus, a series of fixed scenarios must be chosen to evaluate algorithms fairly. We randomly select 10 episodes per month, a total of 120 episodes, which constitute the test dataset. Each episode lasts for 480 time steps (i.e., a day). Then, we split the test dataset into four parts by month: Spring (Mar., Apr., May.), Summer (Jun., Jul., Aug.), Fall (Sept., Oct., Nov.), Winter (Dec., Jan., Feb.). A validation dataset is obtained by randomly selecting 10 episodes from the test dataset. During the training phase, we randomly sample the initial state for an episode and each episode lasts for 240 time steps (i.e., a half day). Each experiment is run with 5 random seeds and the test results during training are given by the median and the 25%-75% quartile shading as same as MAPDN[26]. Moreover, each experiment is evaluated in the validation dataset every 20 episodes during training. After the training phase, we evaluate the learned strategy on the whole test dataset.

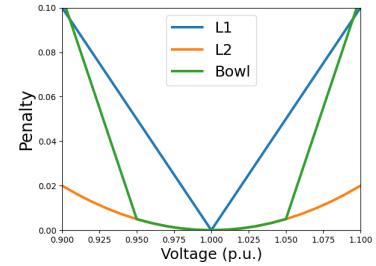
### A.2 Network Topology



**Figure 8: The topologies of power networks. Figures are visualized by the PandaPower toolkit[23].**

The network topologies of the 141-bus network and the 322-bus network are shown in Figure 8. In the 141-bus network, there are 84 loads connected to some specific nodes, and 22 PVs (agents) are installed in some specific nodes. As for the 322-bus network, 337 loads and 38 PVs are connected to some specific nodes.

### A.3 Reward Functions



**Figure 9: Three different voltage barrier functions proposed by [26].**

In this work, reward functions are configured in accordance with the guidelines in MAPDN environment [26]. We also conduct main experiments with different voltage barrier functions using the same settings as [26]: L1-shape, L2-shape and Bowl-shape as shown in Figure 9. Although the action range has been limited, there is still a possibility that the whole power system crashes due to incorrect control actions. To address this issue, if the power system crash, the system would backtrack to the last state and terminate the simulation with a reward of -200 regard as extra penalty.

### A.4 Hyper-parameters of Algorithms.

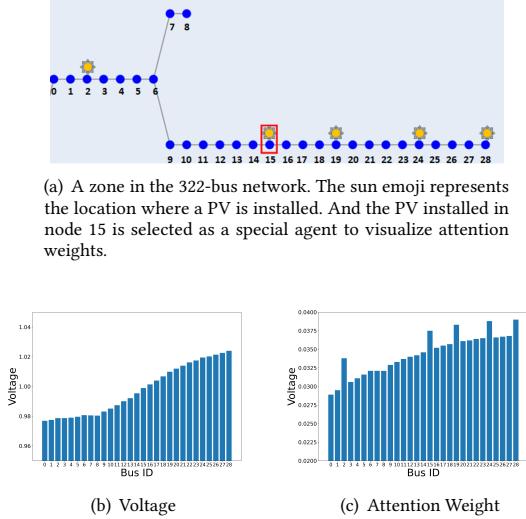
The hyper-parameters in our algorithms are shown in Table 2.

**Table 2: Hyper-parameters in experiments.**

Name	Value
<i>Common:</i>	
optimizer	RMSProp[24]
policy learning rate	$10^{-4}$
value learning rate	$10^{-4}$
policy update epochs	1
value update epochs	10
target update learning rate	0.1
discount factor	0.99
replay buffer size	5000
batch size	32
<i>MADDPG and MATD3:</i>	
MLP hidden dimension (policy)	[64]
GRU hidden dimension (policy)	64
MLP hidden dimension (critic)	[64]
<i>T-MADDPG and T-MATD3:</i>	
auxiliary task learning rate	$10^{-5}$
auxiliary task update epochs	10
transformer hidden dimension (policy)	64
transformer layers (policy)	3
transformer hidden dimension (critic)	64
transformer layers (critic)	3
number of multi-attention heads	4

## B EXTRA EXPERIMENTAL RESULTS

### B.1 Case Study



**Figure 10: A case study on the 322-bus network. We visualize the attention weights of the special agent in (a). The voltages and attention weights of each nodes are shown in (b) and (c), respectively.**

As shown in Figure 10, we visualize the attention weights in the embedding aggregation module to figure out which nodes the agent focuses on. We illustrate a zone in the 322-bus network in Figure 10(a). Five PVs are installed at different locations in the zone. We select the PV installed in node 15 as the special agent, and record attention weights of the final self-attention layer in the embedding aggregation module. As shown in Figure 10(b), the voltage of all nodes is within safety range ( $0.95p.u.$  -  $1.05p.u.$ ). However, due to the radial topology of the distribution network, nodes at the end of the grid face a greater risk of voltage deviations[11]. The trend of attention weights in Figure 10(c) shows that the agent pays more attention to the nodes with high voltage based on the topology of this zone. It is also worthwhile to note that the agent is most concerned with the nodes installed with PVs, which demonstrates that the agent becomes aware of the locations of other PVs in the zone. This phenomenon verifies the opinion discussed in Section 4 that the T-MAAC captures the nature of the gird and extracts better representations.