Enhanced-Visibility Model-Free Deep Reinforcement Learning Algorithm for Voltage Control in Realistic Distribution Systems Using Smart Inverters

Yansong Pei, Ketian Ye, Junbo Zhao, Yiyun Yao, Tong Su, Fei Ding,

***Abstract*—Increasing integration of distributed solar photovoltaic (PV) into distribution networks could result in adverse effects on grid operation. Traditional model-based control algorithms require accurate model information that is difficult to acquire and thus are challenging to implement in practice. This paper proposes a surrogate model-enabled grid visibility scheme to empower deep reinforcement learning (DRL) approach for distribution network voltage regulation using PV inverters with minimal system knowledge. In contrast to existing DRL methods, this paper presents and corroborates the adverse impact of missing load information on DRL performance and, based on this finding, proposes a surrogate model methodology to impute load information utilizing observable data. Additionally, a multi-fidelity neural network is utilized to construct the DRL training environment, chosen for its efficient data utilization and enhanced robustness to data uncertainty. The feasibility and effectiveness of the proposed algorithm are assessed by considering DRL testing across varying degrees of observable load information and diverse training environments on a realistic power system.**

***Index Terms*—Reinforcement learning, active distribution systems, grid visibility, surrogate model, PV inverter.**

I. Introduction

THE increasing penetrations of distributed energy resources (DERs) in distribution systems have reduced the energy burden and environmental impact; however, high penetrations of PVs can have some adverse impacts, such as voltage violations [1]. In recent years, several studies [2], [3] have demonstrated that appropriate real power control and reactive power compensation from smart inverters can impact the PV hosting capacity. The smart inverter achieves voltage regulation by curtailing the active power and/or by supplying or absorbing the reactive power. Although smart inverters have a fast response to voltage violations, the coordination between smart inverters to achieve voltage regulation while reducing active power curtailment needs further consideration.

To control smart inverters, there are many alternative methods based on volt-var control (VVC) [4]–[8]. Although these methods have the advantage of fast response, the predefined volt-var curves are only for each individual smart inverter, and they lack coordinated control. Considering the proliferation of modern power system sizes, designing proper curves is challenging. The optimal power flow-based approaches [9]–[12] can improve system operation by enhancing collaboration across PV inverters. To this end, approximated approaches [10], mixed-integer nonlinear approaches [12], stochastic programming approaches [11], and decentralized approaches [9] have been proposed; however, these approaches assume that system models and parameters are accurate, which is difficult to achieve in practice.

As an alternative, data-driven control techniques using machine learning have been promoted, especially deep reinforcement learning (DRL)-based approaches [13]–[21]. The approach in [13] uses a multi-agent DRL to control switchable capacitors, voltage regulators, and smart inverters to achieve VVC optimization. Reference [14] proposes a soft actor-critic (SAC)-based approach and compares it with traditional VVC to prove that the DRL-based approach works better. A safe off-policy DRL algorithm is proposed for solving the volt-var control problem with voltage-regulating devices in [15]. In [16], a two-timescale voltage control approach is proposed to regulate the voltage in distribution networks by controlling the on-off status of the capacitor units using a deep Q-network. A graph convolution network-based DRL is proposed in [18] to maintain voltage stability in different topologies. In our previous works, [19] proposes a different approach to control PV inverters under changing topology conditions by using a multitask SAC. Although these DRL-based approaches can make control decisions without requiring an accurate system model when training is complete, power flow models or simulators are still required to provide an environment that interacts with the DRL. Once the power flow models or simulators are involved in the training process, the methods cannot be treated as fully data-driven because accurate system parameters and topology are still required. To this end, reference [21] proposes a model-free DRL-based approach with the assumption that all load information is available in real time. This is not practical as only partial loads are measured. In the reference [22], the training environment is conceptualized utilizing a Deep Neural Network (DNN), which serves as a surrogate to the 123-node system. This methodology, while innovative, encounters substantial challenges when applied to real-world, large-scale systems due to the DNN's dependency on extensive datasets of genuine operational data to train and achieve an

This work is supported by the U.S. Department of Energy's Solar Energy Technologies Office Award Number 37770.

Y. Pei, K. Ye, and J. Zhao are with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: yansong.pei@uconn.edu, ketian.ye@uconn.edu, junbo@uconn.edu).

Y. Yao and F. Ding are with National Renewable Energy Laboratory, Golden, CO 80401, USA (e-mail: yiyun.yao@nrel.gov, fei.ding@nrel.gov).

accurate surrogate model. Acquiring such comprehensive data in reality proves to be a formidable task, often limited by practical constraints and the availability of data. Consequently, employing a surrogate model to substitute for large systems as a training environment presents a significant challenge, as the precision of these models is contingent upon the volume and quality of data used for training, which is not readily obtainable in many real-world scenarios.

This paper addresses the aforementioned challenges and develops a visibility-enhanced DRL algorithm for voltage control with only partially observed distribution system load information. The main contributions are:

- Surrogate model-assisted training for reducing the data requirement:The adoption of a multi-fidelity neural network model is proposed to serve as a virtual training environment for DRL algorithms. This model leverages a limited dataset of high-fidelity, real-time operational data complemented by a substantial volume of low-fidelity data.
- Innovative visibility enabled DRL control: Since partial load information will negatively impact DRL control, we construct a visibility enhancement module to infer the unknown load information using partially known load information, voltage information.The estimated load information will be used as an additional DRL observation to improve the control performance.
- SAC-enabled coordinated control: A new reward function design is proposed to balance voltage violation and active power curtailment. Compared to other DRL methods, SAC exhibits superior resilience, especially in the testing phase where it effectively manages data uncertainty in test cases, resulting in notable robustness and proficiency in dealing with uncertain scenarios.

II. Problem Statement

A. Modeling Three-Phase Unbalanced Distribution System

Assume a distribution system has b buses denoted by $\mathcal{B}$:=1, .., b; n nodes denoted by the set $\mathcal{N}$:=1, .., n ; m branches by the set $\mathcal{M}$:=1, .., m.

$$\mathcal{N} = \mathcal{B} \odot \phi, \tag{1}$$

where $\phi = [Phase1, Phase2, Phase3]$ denotes the indicator of whether a bus is a 1-phase or 3-phase bus. For each area/region, there are E nodes denoted by the set $\mathcal{E} \subseteq \mathcal{N}$, which represent that there is an installed voltage sensor. For node $i \in \mathcal{E}$, define v_i^t as the voltage magnitude at t moment. There are H nodes represented by the set $\mathcal{H} \subseteq N$ that have PVs with smart inverters. For $i \in \mathcal{H}$, let the PV set points as $x_i^t := (P_i^t, Q_i^t)$ at time instant t. Assume L nodes denoted by the set $\mathcal{L} \subseteq \mathcal{N}$ have load demand PL_i^t, QL_i^t. Given the restricted access to user load data because of customer premises and hardware considerations, two types of load nodes are anticipated to be present in this system. Let $\dot{PL}_i^t, \dot{QL}_i^t$ the real and reactive power load on node i, which can be observed; there are U nodes denoted by $\mathcal{U} \subseteq \mathcal{L}$ that have unknown load demand; $\hat{PL}_i^t, \hat{QL}_i^t$ are the loads not directly observed. The power flow can be represented by

$$V_i^t (I_i^t)^* = (P_i^t - PL_i^t) + j(Q_i^t - QL_i^t), \forall i \in \mathcal{H}, \tag{2}$$

$$V_i^t (I_i^t)^* = -PL_i^t - jQL_i^t, \forall i \in \mathcal{N}/\mathcal{H}, \tag{3}$$

Throughout the operation of the distribution grid, the nodes equipped with voltage sensors need to be regulated within a predefined secure range. Any nodes of E with higher than 1.05 p.u. or lower than 0.95 p.u. will be counted as voltage violation nodes (VVN). N_{vvn} stands for the total number of VVNs. Formally, we have

$$0.95 \leqslant |V_i^t| \leqslant 1.05, \forall i \in \mathcal{N}/N_{vvn}, \tag{4}$$

$$V_n^t \leqslant 0.95 \ or \ V_n^t \geqslant 1.05, \forall n \in N_{vvn}, \tag{5}$$

B. Coordinated PV Inverter Control for Voltage Regulation

The coordinated PV control aims to optimize specific objectives by regulating the active and reactive power outputs of the PV inverters while respecting the system operation requirements. Fig. 1 illustrates the operational region of PVs. Assume that the PV systems are deployed at nodes $\mathcal{H} \subseteq \mathcal{N}$. The objective of minimizing the PV real power curtailment is

$$f_i^t(p_i^t) = c_{P,i} \times P_i^t, \forall i \in \mathcal{H} \tag{6}$$

where $c_{P,i}$ represents the constant reward coefficient; and P_i^t is the real power generated by the i_{th} PV inverter at time t.

For each PV inverter, the power set point, $rg_i^t := (P_i^t, Q_i^t)$, is constrained to be $rg_i^t \in \mathcal{RE}_i^t$ for $\forall i \in \mathcal{H}$. The feasible region, $\mathcal{RE}_i^t$, is determined by the apparent power capacity and the time-varying solar irradiance, μ_t. Following California Rule 21 on PV interconnection [23], the full extent of the reactive power capability range is defined as 30% of the nameplate apparent power rating. Then, the region $\mathcal{RE}_i^t$ can thus be defined as:

$$\mathcal{RE}_i^t = \{P_i^t + jQ_i^t \mid 0 \leq P_i^t \leq P_{i,max}^t, \tag{7a}$$

$$P_{i,max}^t = \mu_t \times S_i, \tag{7b}$$

$$-0.3S_i \leq Q_i^t \leq 0.3S_i, \tag{7c}$$

$$Q_i^{t\,2} + P_i^{t\,2} \leq S_i^{t\,2}, \} \tag{7d}$$

where $P_{i,max}^t$ is the maximum real power of the i_{th} PV inverter at time t; and S_i is the nameplate apparent power rating of the i_{th} PV inverter.

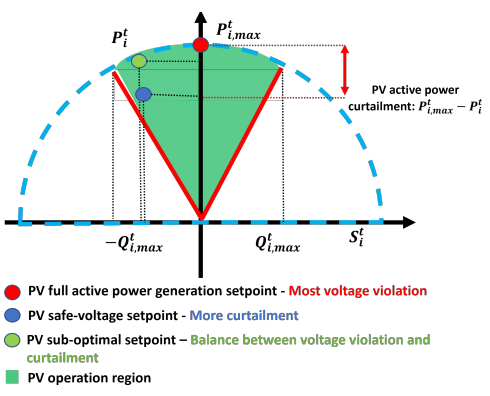

Fig. 1. PV inverter operation regions.

C. Data-Driven Surrogate Models for DRL Training and Execution with Limited Grid Information

This paper aims to build a fully data-driven DRL training and execution framework. It contains two surrogate models, i.e., the observation complementary surrogate model (obs-sur) to enhance grid visibility with partial load information and the training environment surrogate model (train-sur) with a limited number of historical measurements. Considering the important impacts of the missing load information, $\{\hat{PL}_i^t, \hat{QL}_i^t\}$ on DRL decision-making, the purpose of obs-sur is to capture the relationship between the invisible load information and the observable information of the system via $f_{obs} : \boldsymbol{X_{obs}} \rightarrow \boldsymbol{y_I}$. In this surrogate model, X_{obs} represents the visible system information, including the partial active and reactive load power, the voltage profile, and the PV generation information. $\boldsymbol{y_I}$ is the invisible active and reactive power injection of the load. The train-sur is used to provide a virtual environment for the DRL training. This surrogate model is supposed to have the ability to capture the relationship among the load power information, the PV active and reactive power information, and the node voltage profile, which can be represented by the mapping function $f_{train} : \boldsymbol{X} \rightarrow \boldsymbol{Y_v}$. For the training environment surrogate model, $\boldsymbol{X}$ is the load active and reactive power, and the PV active and reactive power set points. $\boldsymbol{Y_v}$ represents all the node voltages that are expected to be regulated.

This paper proposes a multi-fidelity learning framework that allows fusing the low-fidelity model information with a limited number of high-fidelity data for high-fidelity probabilistic voltage predictive analysis. Because both low- and high-fidelity data describe the true system behaviors, they should have correlations, which can be expressed as [24], [25]:

$$\boldsymbol{y}_H = F(\boldsymbol{y}_L) + \delta(\boldsymbol{x}) \tag{8}$$

where $\boldsymbol{x}$ is a vector that denotes the model inputs, such as uncertain PV injections and load; $\boldsymbol{y}_L$ and $\boldsymbol{y}_H$ represent the low- and high-fidelity data, respectively. The low-fidelity data come from inaccurate distribution system model simulations, and numerous data can be generated; the high-fidelity data come from field sensors, such as supervisory control and data acquisition (SCADA) systems and smart meters; $F(\cdot)$ and $\delta(\boldsymbol{x})$ are the nonlinear correlation function and additive correlation term, respectively.

D. Formulation of Markov Decision Process

The coordinated control of the PV active and reactive power set points and the battery energy storage system actions to regulate the voltage and reduce the peak load demand is formulated as an MDP. The MDP comprises the environment, the agents, the observation, the action, and the reward, which are described as follows:

- **Environment**: An active distribution network, including the time-varying load shape. The PV real and reactive power set points will be the input and the output is the voltage of each node, which can be formulated as the following equation:

$$g(P_i^t, Q_i^t, PL_i^t, QL_i^t,) \rightarrow V_m^t, m \in \mathcal{E} \tag{9}$$

In this paper, the load shape and the PV set points will be fed into the pre-trained surrogate model during the training process instead of the simulation software during the testing process, and the voltages are obtained as the output of the surrogate model.

- **Agent**: The central controller of the system. The agent is responsible for controlling the PV inverter set points. In the MDP, the agent makes the decision, A_t, based on the observation, S_t, at the t^{th} time step.
- **Observation**: The information observed by the agent. In this MDP, the agent will observe the time, T, the PV maximum generation, $P_{i,max}^t$, the maximum reactive power capacity, $Q_{i,max}^t$, and the load and EV information, PL_i^t, QL_i^t, consists of load from visible part and prediction data from obs-sur. The set, S_t, including this information, will be used for the agent to make the decision A_t.
- **Action**: The action set, A_t, includes all PV inverter set points. For each PV inverter, $i \in \mathcal{H}$, the action is defined as $(\alpha_{PV,P}(i,t), \alpha_{PV,Q}(i,t))$, where $\alpha_{PV,P}(i,t) \in (0,1)$ and $\alpha_{PV,Q}(i,t) \in (-1,1)$. The PV set points in Eq. (5) can be calculated by the following equation: $P_i^t = \alpha_{PV,P}(i,t) \times P_{i,max}^t, Q_i^t = \alpha_{PV,Q}(i,t) \times 0.3 S_i^t$.
- **Reward**: R_t is obtained after the action, A_t, is executed under the condition of S_t. Considering the different control strategies required for different time periods, two new reward functions are proposed for training:

$$R = -\gamma \sum_{i=1}^{n} v_{i,violation} - \varepsilon Pc^t \tag{10a}$$

$$v_{i,violation} = (1 - min(\delta - |1 - v_i^t|, 0))^2 - 1, \tag{10b}$$

$$Pc^t = 1 - \frac{\sum_{i \in \mathcal{H}} P_i^t}{\sum_{i \in \mathcal{H}} P_{i,max}^t}, \tag{10c}$$

where γ is the penalty coefficient of the voltage violation; ε is the penalty coefficient of the PV active power curtailment according to the PV set points; δ is the threshold used to optimize the voltage barrier function, and Pc_t is the PV active power generation rate used to punish the curtailment.

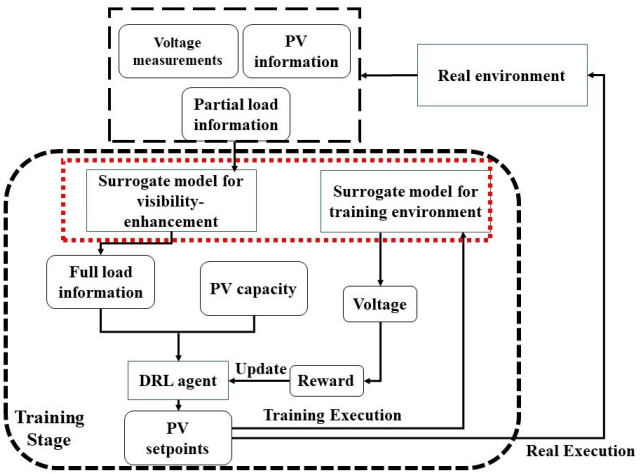

Fig. 2. Overall framework of the proposed scheme.

III. Proposed Grid Visibility Enhanced DRL Control Method

A. Surrogate Model for Data Imputation

With the increasing size of the grid, there are increasing loads in the system. Among the loads, some are not observable (not measured); however, such information is very important because the output of the environment and the decision-making of the DRL largely rely on the full information of the system, especially the power injections and the voltage magnitudes at every node. Missing or insufficient information can have negative effects on the DRL training and execution results [26]. To this end, we propose a surrogate model to impute the missing information, mainly the load injection. This surrogate model with enhanced visibility aims to assist the DRL in the absence of load information. The load and voltage information from other buses is used to infer the missing data:

$$NN_{im} : \{P_k, V_k\} \rightarrow \{P_{uk}, V_{uk}\} \tag{11}$$

where NN_{im} is the neural network for data imputation; P_k and V_k are visible load and voltage information, whereas P_{uk}, V_{uk} are invisible load and voltage to be estimated.

Theoretically, we can derive unknown load values inversely based on power flow equations and known variables, such as load and voltage. Further, the inference of the load shape from load shapes at other buses is possible based on the assumption that loads at geographically close locations have similar patterns or load shapes. The neural networks will learn from both mathematical mapping relationships and load shapes from other buses to produce and impute missing load data. In this paper, deep neural networks are used to impute missing information. It contains multiple layers with each layer fulfilling the nonlinear transformation. Take the l-th layer as an example:

$$NN_{im}^{l} : \delta^{l}(\omega^{l} x + b^{l}) \tag{12}$$

where the weights, ω^t, and bias, b^t, are trainable variables, and δ^t is the activation function. Dropout and skip connections are introduced to enhance the model generalization and to mitigate the gradient diffusion issue. In this paper, the mean square error (MSE) with L-2 regularization is chosen as the loss function:

$$\mathcal{L}_{im} = \sum \mathrm{MSE}(NN_{im}^{l}) + \lambda \sum \omega_i^2 \tag{13}$$

B. Surrogate Model for Training Environment

The environment mentioned in Section II needs to provide the necessary information, including voltage magnitudes, to train the DRL agent for the coordinate control. In most existing DRL-based voltage regulation approaches, the environment is substituted by the simulation software. The environment for the DRL agent can be represented by the mapping function, $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ refer to the uncertain power injections and the voltage magnitudes, respectively; however, the large number of system parameters required by the simulation software to build the environment is contrary to the initial concept of using DRL with little prior knowledge. Based on this assumption, a multi-fidelity learning neural network (NN)-based surrogate model is proposed as a reduced-order model of the simulation software, i.e., $NN : \boldsymbol{x} \rightarrow \boldsymbol{y}$. The multi-fidelity learning framework fuses the low-fidelity model information with high-fidelity data [27]. Because low-fidelity data and high-fidelity data are generated from the same system, a mapping relationship is assumed between them [27]:

$$\boldsymbol{y}_H = \mathcal{F}(\boldsymbol{y}_L, \boldsymbol{x}) \tag{14}$$

where $\boldsymbol{y}_L$ and $\boldsymbol{y}_H$, respectively, represent the low- and high-fidelity data; and $\mathcal{F}$ includes additive, linear, and nonlinear correlations. The low-fidelity data can be massively generated by inaccurate distribution system model simulations, whereas the high-fidelity data come from field sensors, such as SCADA and smart meters.

The multi-fidelity learning framework consists of low-fidelity model construction and high-fidelity model calibration:

$$NN_L : \boldsymbol{x}_L \rightarrow \hat{\boldsymbol{y}}_L, \quad NN_H : \{\boldsymbol{x}_H, \hat{\boldsymbol{y}}_L\} \rightarrow \boldsymbol{y}_H, \tag{15}$$

where $\boldsymbol{x}_L$ and $\boldsymbol{x}_H$ denote the low- and high-fidelity inputs because the input noise is also considered; $\hat{\boldsymbol{y}}_L$ represents the output from the low-fidelity model, and $\boldsymbol{y}_H$ is the high-fidelity output. The multi-fidelity learning structure can be easily extended to T fidelities by stacking correction models hierarchically:

$$\boldsymbol{y}_L^t = \mathcal{F}^t(\boldsymbol{y}_L^{t-1}, \boldsymbol{x}_L^{t-1}), \quad t = 2, \ldots, T \tag{16}$$

where $\mathcal{F}^t$ means the approximate model at t fidelity. Correspondingly, the multi-fidelity neural network is constructed as:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}) = \mathcal{F}^T \circ \mathcal{F}^{T-1} \circ \cdots \mathcal{F}^1(\boldsymbol{x}) \tag{17}$$

The MSE over all fidelity models with L-2 regularization is employed as the loss function:

$$\mathcal{L}_{MF} = \sum_{t=1}^{T} \mathrm{MSE}(NN_{MF}^{t}) + \lambda \sum \omega_i^2 \tag{18}$$

By leveraging the low-fidelity model and the high-fidelity data, the auto-regressive neural networks provide accurate simulation results for the DRL control.

C. SAC-Based DRL Control

DRL is the method of experimenting with different strategies through trial and error to achieve higher rewards. Throughout this process, the neural network of the agent is continuously updated by adjusting the coefficients and weights along the gradients of the high reward. As an advanced DRL approach, actor-critic-based reinforcement learning consists of one actor-network and one critic network. The actor takes the observation as the input and outputs the action accordingly, whereas the critic takes the environment observation along with the actor's action as the input and makes an assessment of the action, providing direction on how to adjust. As the iteration proceeds, the actions given by the actor will receive increasingly higher rewards, and the critic's state value estimation will become more accurate. In contrast to other DRL methods, actor-critic-based methods exhibit rapid convergence and high performance. In this paper, the agent is trained and updated using the off-policy SAC algorithm [28]. The actor-network in the SAC outputs an action by following a policy whose purpose is to maximize the sum of the reward, $R(S_t, A_t)$, and the entropy of the policy, $H(\pi(\cdot \mid s_t))$.

There are three networks in the proposed SAC-based approach: two soft Q-functions networks parameterized by θ and

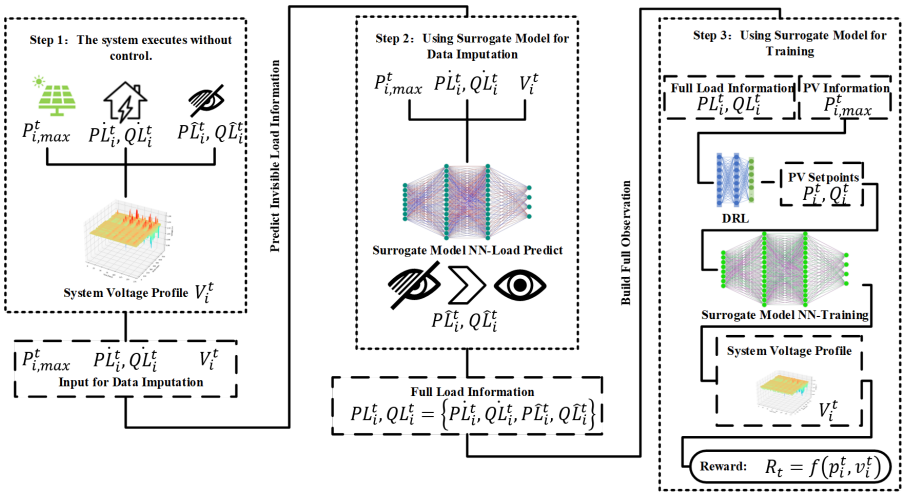

Fig. 3. Proposed two surrogate models enhanced DRL control framework.

a policy function network, π, parameterized by ϕ. The actor-network is a policy equation shown as follows:

$$\mathcal{J}(\pi) = argmax\ E[\sum_{t=0}^{\infty} \omega^t (R(s_t, a_t) + \alpha \times H(\pi(\cdot \mid s_t))], \quad (19)$$

where ω is the future discount coefficient; and α is a temperature parameter that indicates the entropy's contribution to the reward. α will initially be designed as a large value to obtain higher entropy rewards by increasing the exploration space of action. As the training proceeds, α will gradually decrease and shrink the exploration breadth, eventually approaching the optimal policy. The critic network estimates the state value of the Q-function as:

$$y(S_t, R_t, S_{t+1}) = r + \omega(Q(S_{t+1}, A_{t+1} - \alpha log \pi_\theta(A_{t+1} \mid S_{t+1})), \quad (20)$$

The SAC algorithm relies on an experience replay buffer to update with enhancing sample efficiency. After the reward is obtained by the executed action, the replay buffer stores the observation, the action, the reward, and the next step observation as a transition. A batch of transitions, $B = \{(S_t, A_t, S_{t+1}, R_t)\}$, will be randomly selected to update the neural network. The actor-network updates the coefficient using gradient ascent by the following:

$$\nabla \phi_i \frac{1}{|B|} \sum_{((s_t)) \in B} (\min_{i=1,2} Q_{\theta_i}(s_t, \pi(\cdot \mid s_t)) - \alpha \log \pi_\theta(\pi(\cdot \mid s_t) \mid s_t)), \quad (21)$$

The critic network updates the Q-function using gradient descent by following:

$$\nabla \theta_i \frac{1}{|B|} \sum_{((s_t, a_t, s_{t+1}, r_t)) \in B} (Q_{\theta_i}(s_t, a_t) - y(s_t, r_t, s_{t+1}))^2, \quad (22)$$

where the clipped double-Q method is used to obtain the smaller Q-value between the two Q-approximators. Finally, the gradient rule is applied to update the actor-network by following:

$$\phi_{target,i} \leftarrow \rho \phi_{target,i} + (1 - \rho)\phi_i \quad (23)$$

where ρ represents the learning rate for the actor-network. The training process of coordinated control using DRL is implemented via Algorithm 1.

Algorithm 1 Training process of coordinated PV inverter control

```
Initialization all agent NN parameters φ_i, θ_i
Initialization episode reward list list_eps
for episode = 1 to eps do
    Initialization episode reward R_eps = 0
    for step = 1 to t do
        Get S_t = {PL_i^t, QL_i^t, P_{i,max}^t, Q_{i,max}^t}
        if episode ≤ M then
            Randomly generate A_t
        else
            Input S_t to agent and Output A_t
        end if
        Get α_{PV,P}(j,t), α_{PV,Q}(j,t) based on A_t;
        Execute x_t in train-sur model
        Get v_t, x_t to calculate R_t
        Let R_eps = R_eps + R_t
        Store {s_t, A_t, s_{t+1}, R_t} in replay buffer
        Randomly sample batch B = {(s_t, a_t, s_{t+1}, r_t)}
        Update crtic network θ_i by following (22)
        Update actor network φ_i by following (21) (23)
    end for
    list_eps append(R_eps)
    if Max(list_eps) = R_eps then
        Store the agent NN parameters φ_i, θ_i
    end if
end for
```

D. Surrogate Model-Assisted Training and Execution

The proposed method includes a parameter set required to be optimized as $\Psi = \{NN_{im}, \mathcal{F}(\boldsymbol{x}), \phi_i, \theta_i\}$, where $\{NN_{im}$ is the parameter of obs-sur, $\mathcal{F}(\boldsymbol{x})$ is the parameter of the train-sur, and ϕ_i, θ_i are the parameters of the DRL agent. The surrogate model for the data imputation obs-sur and the surrogate model for the training environment train-sur will use

the historical data to complete the training in advance. Then, they will be integrated into the proposed framework, which can assist the DRL in training with visibility to enhance using two surrogate models. The DRL training procedures are include:

1) **Visibility enhancement using obs-sur**: The obs-sur is first trained using historical data before the DRL training. Once the obs-sur training is finished, the parameters of obs-sur, $NNim$, will be stored. After the training begins, $NNim$ is downloaded and loaded into the local network. At the beginning of each time step, the system will be operated without any control of PV. The observed information—including $\dot{PL}_i^t$, $\dot{QL}_i^t$, P_i^t, Q_i^t, V_i^t—will be used to predict the unknown load information, $\hat{PL}_i^t, \hat{QL}_i^t$. The observed load information, $\dot{PL}_i^t$, $\dot{QL}_i^t$, and the predicted load information, $pred\hat{PL}_i^t, \hat{QL}_i^t$, will form a set of PL_i^t, QL_i^t used to help the DRL make decisions and apply control strategies.

2) **DRL training using train-sur**: After the full visible load information, PL_i^t, QL_i^t, is obtained, the observation set, S_t = PL_i^t, QL_i^t, $P_{i,max}^t$, $Q_{i,max}^t$, will be sent to the DRL agent to make action A_t. Most existing approaches rely on an accurate model to build a virtual power system in the simulation software. Unlike the simulation software, our proposed surrogate model needs only a few high-fidelity data and low-fidelity data. These data can be easily obtained from the historical data set compared with the model information and line parameters. Compared with existing approaches, the action, A_t, will be executed in the environment provided by train-sur instead of the simulation software. The surrogate model will take A_t, S_t as the input and the voltages as the output. The voltage and action A_t will be used to calculate the reward based on the specially designed reward function and update the DRL agent by following **Algorithm 1**. During this process, the threshold, δ, mentioned in Section II can also be used to offset the errors generated during the surrogate model calculation of the voltage profile.

3) **Implementation of the proposed method**: When the DRL agent training process is completed, the parameters of the actor neural network, ϕ_i, and obs-sur neural network, NN_{im}, will be stored locally for execution. Only the obs-sur and the actor-network of the DRL agent will be involved in the execution process for the purpose of providing full observation and control solutions. The algorithm implementation is shown in **Algorithm 2**.

In **Algorithm 2**, at the beginning of each time step, there is no control command given to the PV. The PV inverter will generate the maximum amount of active power during the daytime with solar irradiation and no active power during the nighttime. During this process, no reactive power will be generated or absorbed. The voltage profile of the system will be obtained after the execution of the PV set points. Then, the set of $\{\dot{PL}_i^t, \dot{QL}_i^t, P_{i,max}^t\}$ will be fed into the obs-sur to predict the unobservable load information. The output of the obs-sur $\{\dot{PL}_i^t, \dot{QL}_i^t\}$ combined with the known load data is considered full load information. The maximum PV active power generation with full load information will be treated as the observation of the DRL agent. The DRL agent will follow the policy A_t=$\pi(s_t)$ to make the action decision. The PV set points in the action set will be sent to the PV inverters and executed. The neural networks in the surrogate mode are very fast in computation and operation, thus guaranteeing a real-time response.

Algorithm 2 Execution process of coordinated PV inverter control

```
Create the neural networks of the DRL agent and obs-sur
Load the parameter φ_i in the neural network of the DRL
Load the parameter NN_im in the neural network of obs-sur
for step = 1 to t do
    if P^t_{i,max} ≠ 0 then
        PV execute P^t_{i,max}
    else
        PV has no action
    end if
    Obtain voltage profile, v_i^t
    Input {P̀L_i^t, Q̀L_i^t, P^t_{i,max}, v_i^t} in obs-sur with the parameter NN_im
    Obs-sur predicts the invisible load {P̂L_i^t, Q̂L_i^t}
    let {PL_i^t, QL_i^t} = {P̀L_i^t, Q̀L_i^t, P̂L_i^t, Q̂L_i^t}
    Get S_t = {PL_i^t, QL_i^t, P^t_{i,max}, Q^t_{i,max}}
    DRL actor network Output A_t by following A_t=π(s_t)
    Get x_t = α_{PV,P}(j,t), α_{PV,Q}(j,t) based on A_t
    PV execute x_t in system
end for
```

IV. Testing Results

The comparative experiments are conducted using a realistic 759-node model located in Western Colorado, as illustrated in Fig. 4. This system comprises a total of 623 buses, among which 68 buses are equipped with 3-phase capability. We compare our method with the traditional VVC by following the voltage-reactive power curve in [23]. Additionally, we employ a different visibility DRL agent for comparison, to demonstrate the effectiveness of our framework. For the test system, a total of 112 PV units are used for active power generation and the exchange of reactive power to regulate the voltage in the system. The system includes a total of 159 loads; 59 are observable, and the remaining 100 are unobservable. The peak load on the system is 2.61 MW, and the installed PV systems have a capacity of 1.12 MW; the PV systems are integrated into the power grid at a level of 43% penetration. The substation will be responsible for supplying the remaining electricity. To train the two surrogate models and the DRL agent, the solar radiance data and the historical data pertaining to load and voltage are used. The aforementioned data are divided into a training set and a test set, respectively, which are recorded at 5-minute intervals. The surrogate model for the data imputation is trained using 5×10^3 sets of historical data that include the input data, $\dot{PL}_i^t, \dot{QL}_i^t, P_{i,max}^t, v_i^t$, and the output data, $\hat{PL}_i^t, \hat{QL}_i^t$. The unobservable load is considered not immediately visible, but it can be obtained through calculations and retroactive analysis of historical data. The parameters of NN_{im} are obtained and stored. The surrogate model for the training environment is

trained using 10^3 high-fidelity data collected during accurate model operation and 10^5 low-fidelity data generated by an inaccurate model in the simulation software. High-fidelity data refers to OpenDSS simulation results without model error while low-fidelity data is obtained from OpenDSS simulation with Gaussian noises on both input and output. The input noise is Gaussian noise with a zero mean and a standard deviation equals to 20% of the original load values. Similarly, the output noise follows a Gaussian distribution with a mean of zero and a standard deviation of 0.01% of the voltage values. The imprecise load shape information, combined with the solar irradiance data gathered from field measurements in Denver, $PL_i^t, QL_i^t, P_{i,max}^t, Q_{i,max}^t$, will be input into the simulation software. In this process, the voltage profile, v_i^t, is calculated by following the power flow model in the software under the condition that the PV units operate with random set points. The collection of the data mentioned here will constitute the low-fidelity data set and will be used to update the train-sur. The parameters of the train-sur will be stored for the DRL training. The hyperparameters are shown in Table.I. OpenDSS is used to execute the power flow, and the training of the proposed DRL method is implemented in Python with PyTorch. A workstation with an NVIDIA GeForce RTX 3090 GPU and Intel Core i9-12900KF is used for the simulation.

TABLE I
DRL PARAMETER SETTING

Parameters	Values
Batch size for updating NN	32
Replay buffer size	48000
Discount factor	0
Soft update coefficient	0.001
Target policy smoothing coefficient	0.2
Learning rate for actor network	0.001
Learning rate for critic network	0.002

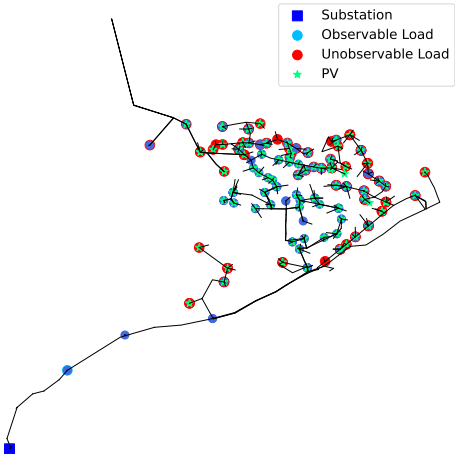

Fig. 4. The realistic 759-node distribution system in western Colorado.

A. Reward Function Configuration and Pre-training

The design of the reward function is crucial for the successful training of DRL. The trade-off between minimizing voltage violations and minimizing active power curtailment is inherent. When the value of γ is set too high, DRL will focus on voltage control, leading to a significant amount of active power curtailment. Conversely, if ϵ is set too high, the DRL will prioritize increasing PV generation, which may result in increased voltage violations.

TABLE II
REWARD TESTING

ratio of γ and ϵ	VVN	Curtailment
40	13	11.6%
20	13	5.68%
10	22	2.13%
1	24	1.84%

Therefore, γ and ϵ as individual parameters do not have physical significance. Instead, the ratio between γ and ϵ has a greater impact on DRL training. The primary objective of this work is to ensure voltage stability in the power system, followed by the reduction of PV curtailment. Based on this principle, we trained normal SAC using different reward function using 1-day data and test result is shown in Table.II. we observed that when the ratio is set too high, beyond a certain threshold, further increases in curtailment do not lead to a reduction in violations. Consequently, we selected a final ratio of 20:1 for training both the proposed method and the comparison method.

To further demonstrate the stability and replicability of our training environment, we conducted multiple training sessions using the same surrogate model across various training platforms. The results are illustrated in the Fig.5. The training processes of four agents are depicted. The training process includes both exploration and exploitation phases. Before 400 steps, due to different random seeds and learning processes, the episode rewards during the exploration phase vary. However, as the training episodes increase, the rewards of all four agents gradually converge to the same range. Therefore, we believe that as long as the hyper-parameters and training environment are consistent, our proposed method is replicable.

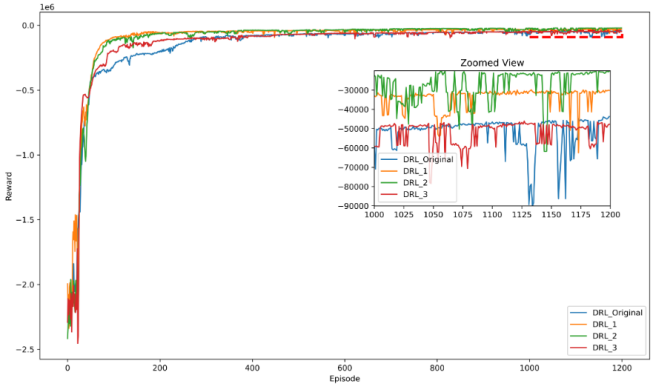

Fig. 5. DRL Agents Training Process

B. Performance Evaluation of the Visibility-Enhancement Surrogate Model

The obs-sur that has been trained is implemented during both the training and testing stages. During the testing stage, the data of load and solar irradiance are collected at a 1-hour time resolution, resulting in a total of 24 steps. Fig. 6 show comparisons between the actual load data and the predicted load information using obs-sur in the test process. The trend of the total load variation during the testing phase, comprising both real and predicted data, is illustrated in Fig. 6. The horizontal axis represents time in hours, ranging from 0 to 24 hours, which indicates a full day's cycle. The vertical axis on the left shows the Load Power in kilowatts (kW) for individual predictions and real values, while the vertical axis on the right represents the Total Power in kilowatt-hours (kWh), demonstrating the cumulative load.

In the 24-hour prediction window, the actual data exhibited a peak load of 32.048 kW, while the surrogate model predicted a slightly higher load of 32.810 kW, resulting in an error margin of 2.38%. Throughout these 24 steps, the real total load amounted to 5103 kWh, and the predicted total load was 5135 kWh, yielding a minor discrepancy of 0.64%. The overall mean absolute error (MAE) for the prediction stood at 0.092. These figures demonstrate that the obs-sur model provides highly accurate predictions, implying that it offers precise observational data for decision-making in DRL frameworks.

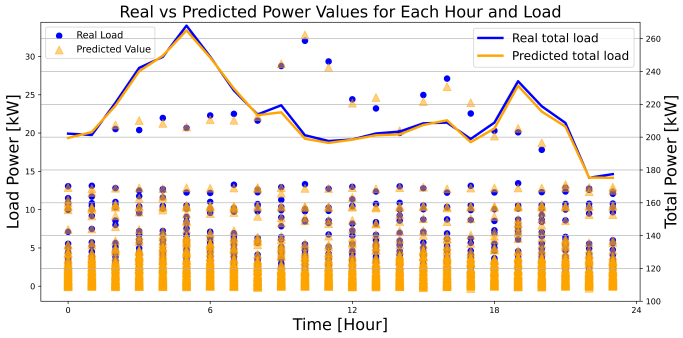

Fig. 6. Total load prediction in the test process.

TABLE III
LOAD PREDICTION INFORMATION

Data	Time step	Max [kW]	Min [kW]	MAE	Total load [kWh]
Real data	24	32.048	0	0.092	5103.2
Obs-sur		32.810	-0.303		5135.8

C. Performance Evaluation of the Training Environment Surrogate Model

To better demonstrate the efficiency of multi-fidelity networks in data utilization, a control group was established to highlight the advantage of the networks. The terms of LF, HF, and HLF represent training scenarios involving large amounts of low fidelity data, small quantities of high fidelity data, and a combination of both, respectively, for conventional

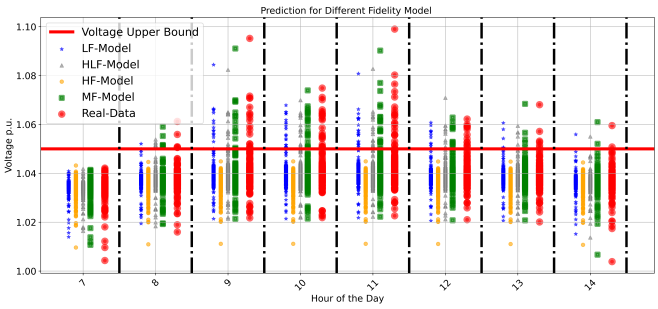

Fig. 7. Surrogate Model Prediction Voltage Comparison

neural network training. The MF model indicates the use of a MF neural network with the same data as the HLF model. The testing phase was based on baseline scenario, involving 112 PV inverters where the active power fluctuates with changes in sunlight and reactive power remains zero, and 159 load nodes with both active and reactive power inputs. This data was used to predict the voltage at 759 nodes using a surrogate model, over a 24-hour period with hourly time steps. Voltage violation issues were only observed when solar intensity reached a certain threshold, thus Fig.7 shows the voltage prediction result from 9AM to 4PM. Considering that most node voltages remain within normal ranges at all times, the graph under the 1.05 p.u. voltage threshold (red line) is stacked. Table.IV presents a comparison in terms of the mean absolute error (MAE), maximum and minimum voltages, and the number of voltage violations. Models using limited high fidelity data performed the worst, failing to provide reliable voltage violation information. In contrast, the HLF model yielded more accurate predictions with an MAE reduced to 0.0011, but due to the direct combination of high and low fidelity data, the HLF model sometimes under performed compared to the LF model. From the perspective of MAE, the MF-model slightly outperforms other models. This is primarily because the voltages of the 759 nodes are largely concentrated, leading to a dilution of the prediction error for specific nodes. Additional details are evident in Fig.7, particularly at 11 AM and 1 PM. Although all models exhibit some level of error, the MF model (represented by green squares) aligns most closely with the actual values (indicated by red circles). Moreover, at other times, the MF model provides predictions that are closest to the actual voltage distribution, including the most accurate forecasts of maximum and minimum voltage values. Consequently, compared to other models, MF demonstrates superior data utilization efficiency and offers a more accurate predictive environment for subsequent DRL training.

TABLE IV
VOLTAGE PREDICTION INFORMATION

Data	Max p.u.	Min p.u.	MAE	VVN
Real	1.099	0.976	-	95
HF-Model	1.045	0.975	0.0022	0
HLF-Model	1.083	0.975	0.0011	78
LF-Model	1.084	0.975	0.0010	80
MF-Model	1.091	0.975	0.0005	82

D. Voltage Control Performance Evaluation

Upon completing the training process employing two surrogate models, a comparative analysis is conducted using 1 day's data in a 1-hour resolution to assess the performance of the proposed methodology. The benchmark methods evaluated in this comparison encompass: **1) Non-control**: All PV units generate as much active power as possible, and no reactive power will be produced or absorbed. **2) Autonomous volt-var control**: Each PV unit will deliver a reactive power response adhering to the Category A curve as delineated in the IEEE 1547-2018 standard [23]. A deadband is incorporated with the voltage magnitude range defined as [0.98,1.02]. Two distinct configurations are employed: VVC-watt priority and VVC-var priority. In the VVC-watt priority scheme, the reactive power set points are set to be the value requested by the volt-var curve, which is limited by each inverter's capacity. Conversely, in the VVC-var priority approach, the active power generation is curtailed until the desired reactive power set points are attained; **3) DDPG-Full & DDPG-Sur**: Both the full observation deep deterministic policy gradient (DDPG) DRL agent and its low observation counterpart, DDPG-Sur, are trained using a precise model in simulation software and an obs-sur environment, respectively. During training and testing, they can access the active and reactive power setpoints for all 159 load nodes. **4) SAC-Full**: The SAC-based DRL agent with full observation is trained in a simulation software that employs a precise model. Throughout the training and testing phases, information including 159 load active and reactive power set points can be observed. **5) SAC-Low**: The SAC-based DRL agent with low observation is trained in simulation software that employs a precise model. Throughout the training and testing phases, information including 59 load active and reactive power set points can be observed. **6) SAC-Surrogate**: The SAC-based agent, enhanced with low observation capabilities in the obs-sur environment, is provided with the active and reactive power setpoints for all 159 load nodes. Subsequent training is conducted within surrogate model environments as LF, HF, and HLF, and the agents trained in these respective environments are denoted as SAC-LF, SAC-HF, and SAC-HLF. Table V includes the experimental results and the settings for the training environment and observation. The methods of non-control, volt-var control, and surrogate model-assisted training do not rely on a real simulation environment for training. Similarly, the observation settings are divided into "full" and "part," representing the ability to observe the entire system's load conditions and the ability to observe only part of the load conditions, respectively. Table V shows that the absence of control will lead to serious voltage violations, as indicated by the maximum voltage of 1.082 p.u. During the entire test, a total of 71 voltage violation nodes are recorded, and the use of VVC-watt priority is proved to be of limited effectiveness in voltage control. This is because VVC-watt prioritizes the regulation of reactive power while ensuring a steady output of active power, thus leaving little capacity for voltage regulation. As a result, the maximum voltage is reduced to 1.063 p.u., leading to a decrease in the total number of violation nodes to 30. VVC-var priority, which adjusts the reactive power to the desired level before generating the active power, further decreases the maximum voltage violation to 1.059 at curtailment of 3.45%. Despite this improvement, 18 voltage violation nodes remain unresolved.

TABLE V
COMPARISON RESULT FOR DIFFERENT APPROACHES.

Method	Max p.u.	Min p.u.	VVN	Curt	env	obs
Non-control	1.082	0.976	71	-	✗	-
VVC-watt	1.063	0.976	30	-	✗	**Full**
VVC-var	1.059	0.975	18	3.45%	✗	**Full**
DDPG-Full	1.062	0.961	7	34.6%	✓	**Full**
DDPG-Sur	1.064	0.962	5	43.1%	✗	**Part**
SAC-Full	**1.051**	**0.963**	**1**	**5.42%**	✓	**Full**
SAC-Low	1.062	0.975	21	5.73%	✓	Part
SAC-LF	1.066	0.963	11	3.83%	✗	Part
SAC-HF	1.088	0.976	82	0.46%	✗	Part
SAC-HLF	1.066	0.966	11	3.88%	✗	Part
Proposed	**1.055**	**0.972**	**6**	**4.26%**	✗	**Part**

In contrast to traditional VVC methods, DDPG-Full and DDPG-Sur succeeded in reducing the number of VVN to 7 and 5, respectively. However, achieving these outcomes required substantial curtailments of 34.6% and 43.1%, which are significantly high. The approach of using SAC trained in a real environment, with complete load observation, yields substantial improvements in voltage control. Specifically, the maximum voltage violation has been reduced to 1.051 p.u., and the total number of violations is reduced to 1 with 5.42% curtailment. Note that when the number of observable loads is decreased, the performance of SAC-Low is poor, leading to a maximum voltage violation of 1.062 p.u. and a total of 21 violation nodes. This decline in performance is accompanied by an increased curtailment of 5.73%. When the training environment transitions from real simulation software to surrogate models, the result will be influenced by different surrogate models. Due to the HF-model employing only a minimal amount of high-fidelity data, it is evident from Table.IV that the HF-model is incapable of predicting any voltage violations. Consequently, within the HF-model environment, the DRL disregards voltage violation issues, aiming to minimize curtailment as much as possibile, resulting in 82 VVNs. In contrast, both the LF-model and the HLF-model provide similar voltage prediction training environment, leading to 11 occurrences of voltage violations during testing, with curtailment rates of 3.83% and 3.88%, respectively. However, when training DRL with the proposed MF-model, the performance of the proposed SAC and the SAC trained in the real environment (SAC-full) are most aligned, exhibiting only six voltage violation points and a curtailment rate of 4.26%. Considering that the proposed method does not utilize any real-world data and relies entirely on information provided by two surrogate models, this performance is exceptionally commendable. The Table.VI provides voltage statistics across different phases. It is observed that in the non-controlled baseline scenario, voltage violations primarily occur in phases 2 and 3, with phase 3 experiencing the most severe violations. In contrast, the SAC-Full approach, which is informed by real-world environments and complete

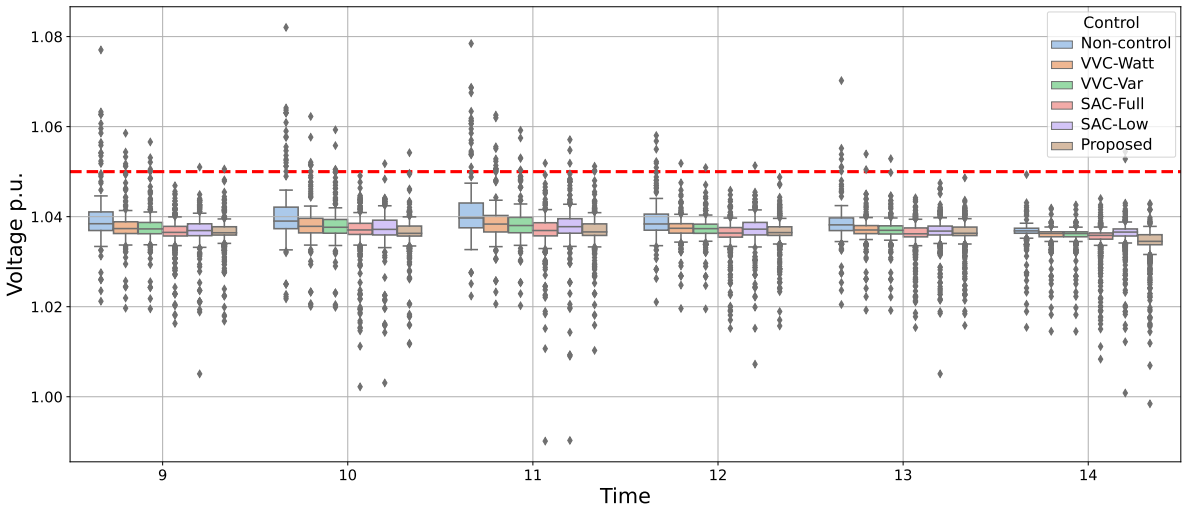

Fig. 8. Voltage comparison results during the test process.

TABLE VI
COMPARISON OF VOLTAGE FOR DIFFERENT PHASES

		Phase 1	Phase 2	Phase 3
Baseline	Max	1.039	1.068	1.082
	Min	1.031	0.976	0.991
SAC-Full	Max	1.049	1.051	1.044
	Min	1.013	0.974	0.963
Proposed	Max	1.045	1.055	1.053
	Min	1.014	0.974	0.972

observational data, completely eliminates voltage violations in phase 3, with only a minor violation of 1.051 p.u. remaining in phase 2. The proposed method, despite operating under the least favorable conditions with model-free training, manages to control the maximum voltage violations in phases 2 and 3 to 1.055 and 1.053 p.u., respectively.

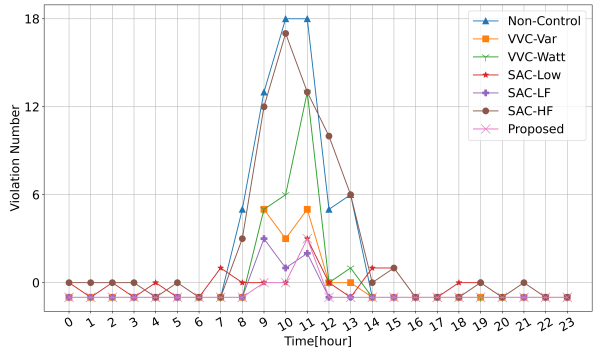

Fig. 9. The number of VVNs for different approaches.

The trend and the comparative analysis of the effectiveness of different control methods during the test day are illustrated in Fig. 9. Specifically, the SAC-Full demonstrated an almost impeccable adherence to voltage limits, with a singular incident of violation observed over the 24-hour periods. In contrast, the SAC-Low showed a more varied pattern of breaches, albeit relatively minor, with occurrences ranging from one to five violation in a few periods, highlighting its less consistent control over voltage thresholds. The proposed approach with violation recorded in only three steps, and even then, the instances were limited to a maximum of four violation in the most challenging period. This starkly contrasts with the non-control situation, which experienced a significant surge in violation, particularly in the middle of the day, where the numbers soared to as high as 18. The traditional VVC-Var and VVC-Watt methods exhibited intermediate performance. The VVC-Var method demonstrated a moderate level of control, with violation peaking at six in certain periods but otherwise maintaining a low profile. Similarly, the VVC-Watt method showed a balanced performance, with its violation numbers mirroring those of VVC-Var in some periods but also peaking at 14 in a single period, indicating occasional challenges in voltage regulation.

Fig. 8 displays the voltage distribution at each hour from 9 a.m.–2 p.m. At 10 a.m., the non-control results in the highest voltage level, exceeding 1.08 p.u., whereas all other control methods effectively reduce the voltage to less than 1.07 p.u. Notably, our proposed method demonstrates superior performance, with the highest voltage level recorded at 1.055 p.u.; however, SAC-Low exhibits instability starting at 10 a.m. As solar radiation intensity decreases over time, the maximum voltage also gradually decreases, with no voltage violations observed after 2 p.m. except for in the SAC-Low. Although the use of large curtailment by DDPG to reduce the number of VVNs is not ideal, the surrogate model provided the agents with a training environment highly similar to the real-world scenario, leading to very similar performance and results between DDPG-Full and DDPG-Sur. Overall, SAC-Full and the proposed approach exhibit the best performance, achieving a commendable balance between reducing curtailment and controlling voltage. In some extreme cases, they also demonstrate superior adaptability compared to traditional methods, further meeting the requirements for voltage control. In terms of operational efficiency, our system includes two key components: the obs-sur and the train-sur, along with the DRL agent. The obs-sur is responsible for predicting unobservable

values from observable ones and providing these predictions to the DRL agent for decision-making. Our tests indicate that this process requires an average of 0.003 seconds per step. Conversely, the train-sur component serves as the training environment for the DRL agent. It significantly enhances efficiency with a neural network output time of just 0.00087 seconds per step, which is substantially faster compared to the 0.013 seconds per step required by OpenDSS. During the execution phase, the DRL agent makes decisions within 0.001 seconds after receiving the state information from obs-sur. Combining this with the processing time of obs-sur, the total decision-making duration is kept under 0.004 seconds. Given these results, we assert that the efficiency and speed of our proposed method qualify it for real-time applications. The proposed approach, as a purely data-driven method, is better suited to meet the needs of real-world power systems.

V. Conclusion

In this paper, a new surrogate model-enhanced DRL control framework is proposed. Initially, the neural network is established to map the relationships among the voltage, the visible load, and the invisible load, and it is used to predict the behavior of the invisible loads. Subsequently, we leverage the data efficiency of multi-fidelity neural networks to create a training environment for DRL that requires reduced amounts of accurate data. The proposed framework ultimately implements a data-driven DRL mode that offers a promising approach to address the challenges associated with voltage control. Comparative results demonstrate that: 1) DRL without sufficient observation will degrade performance and stability. 2) In the proposed multi-fidelity surrogate model, DRL trained in a high-precision environment can achieve performance results similar to those trained in real-world environments. 3) Surrogate model-assisted DRL achieves the best control performances as compared to other methods. Future work will focus on enhancing the training efficiency of the surrogate model to handle larger-scale distributed systems. Additionally, when dealing with vast systems, we plan to regionalize the system and use multiple surrogate model-DRL agent configurations to ensure the scalability of our proposed method.

References

[1] S. Eftekharnejad, V. Vittal, G. T. Heydt, B. Keel, and J. Loehr, "Impact of increased penetration of photovoltaic generation on power systems," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 893–901, 2013.

[2] J. Seuss, M. J. Reno, R. J. Broderick, and S. Grijalva, "Improving distribution network pv hosting capacity via smart inverter reactive power support," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2015, pp. 1–5.

[3] F. Ding, B. Mather, and P. Gotseff, "Technologies to increase pv hosting capacity in distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2016, pp. 1–5.

[4] S. Yoshizawa, Y. Yanagiya, H. Ishii, Y. Hayashi, T. Matsuura, H. Hamada, and K. Mori, "Voltage-sensitivity-based volt-var-watt settings of smart inverters for mitigating voltage rise in distribution systems," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 584–595, 2021.

[5] H. Lee, J.-C. Kim, and S.-M. Cho, "Optimal volt–var curve setting of a smart inverter for improving its performance in a distribution system," *IEEE Access*, vol. 8, pp. 157 931–157 945, 2020.

[6] R. A. Jabr, "Linear decision rules for control of reactive power by distributed photovoltaic generators," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2165–2174, 2018.

[7] M. Jalali, V. Kekatos, N. Gatsis, and D. Deka, "Designing reactive power control rules for smart inverters using support vector machines," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1759–1770, 2020.

[8] Y. Chai, L. Guo, C. Wang, Z. Zhao, X. Du, and J. Pan, "Network partition and voltage coordination control for distribution networks with high penetration of distributed pv units," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3396–3407, 2018.

[9] Y. Yao, F. Ding, K. Horowitz, and A. Jain, "Coordinated inverter control to increase dynamic pv hosting capacity: A real-time optimal power flow approach," *IEEE Syst. J.*, pp. 1–12, 2021.

[10] B. Stott, J. Jardim, and O. Alsac, "Dc power flow revisited," *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1290–1300, 2009.

[11] Y. Xu, Z. Y. Dong, R. Zhang, and D. J. Hill, "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4398–4408, 2017.

[12] Y. J. Kim, J. L. Kirtley, and L. K. Norford, "Reactive power ancillary service of synchronous dgs in coordination with voltage control devices," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 515–527, 2017.

[13] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, 2021.

[14] Y. Pei, Y. Yao, J. Zhao, F. Ding, and K. Ye, "Data-driven distribution system coordinated pv inverter control using deep reinforcement learning," in *2021 IEEE Sustainable Power and Energy Conference (iSPEC)*, 2021, pp. 781–786.

[15] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, 2020.

[16] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2313–2323, 2020.

[17] J. Duan, D. Shi, R. Diao, H. Li, Z. Wang, B. Zhang, D. Bian, and Z. Yi, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, 2020.

[18] R. R. Hossain, Q. Huang, and R. Huang, "Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4848–4851, 2021.

[19] Y. Pei, J. Zhao, Y. Yao, and F. Ding, "Multi-task reinforcement learning for distribution system voltage control with topology changes," *IEEE Trans. Smart Grid*, pp. 1–1, 2022.

[20] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention enabled multi-agent drl for decentralized volt-var control of active distribution system using pv inverters and svcs," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1582–1592, 2021.

[21] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, Z. Chen, and F. Blaabjerg, "Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of pvs," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4137–4150, 2021.

[22] D. Cao, J. Zhao, W. Hu, F. Ding, N. Yu, Q. Huang, and Z. Chen, "Model-free voltage control of active distribution system with pvs using surrogate model-based deep reinforcement learning," *Applied Energy*, vol. 306, p. 117982, 2022.

[23] California Public Utilities Commission, "Rule 21 interconnection." [Online]. Available: https://www.cpuc.ca.gov/Rule21

[24] M. Giselle Fernández-Godino, C. Park, N. H. Kim, and R. T. Haftka, "Issues in deciding whether to use multifidelity surrogates," *AIAA Journal*, vol. 57, no. 5, pp. 2039–2054, 2019.

[25] B. Peherstorfer, K. Willcox, and M. Gunzburger, "Survey of multifidelity methods in uncertainty propagation, inference, and optimization," *Siam Review*, vol. 60, no. 3, pp. 550–591, 2018.

[26] T. Katakura, M. Yoshida, H. Hisano, H. Mushiake, and K. Sakamoto, "Reinforcement learning model with dynamic state space tested on target search tasks for monkeys: Self-determination of previous states based on experience saturation and decision uniqueness," *Frontiers in Computational Neuroscience*, p. 134, 2022.

[27] K. Ye, Y. Pei, J. Zhao, Y. Yao, J. Wang, and F. Ding, "Multi-fidelity learning for distribution system voltage probabilistic analysis with high penetration of pvs," in *Proc. IEEE Power Energy Soc. General Meeting*, 2022, pp. 1–5.

[28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018, pp. 1861–1870.