Active Distribution System Synthesis via Unbalanced Graph Generative Adversarial Network

Rong Yan, *Graduate Student Member, IEEE*, Yuxuan Yuan, *Graduate Student Member, IEEE*, Zhaoyu Wang, *Senior Member, IEEE*, Guangchao Geng, *Senior Member, IEEE*, and Quanyuan Jiang, *Senior Member, IEEE*

***Abstract*—Real active distribution networks with associated smart meter (SM) data are critical for power researchers. However, it is practically difficult for researchers to obtain such comprehensive datasets from utilities due to privacy concerns. To bridge this gap, an implicit generative model with Wasserstein GAN objectives, namely unbalanced graph generative adversarial network (UG-GAN), is designed to generate synthetic three-phase unbalanced active distribution system connectivity. The basic idea is to learn the distribution of random walks both over a real-world system and across each phase of line segments, capturing the underlying local properties of an individual real-world distribution network and generating specific synthetic networks accordingly. Then, to create a comprehensive synthetic test case, a network correction and extension process is proposed to obtain time-series nodal demands and standard distribution grid components with realistic parameters, including distributed energy resources (DERs) and capacitor banks. A Midwest distribution system with 1-year SM data has been utilized to validate the performance of our method. Case studies with several power applications demonstrate that synthetic active networks generated by the proposed framework can mimic almost all features of real-world networks while avoiding the disclosure of confidential information.**

***Index Terms*—Graph generative adversarial network, network synthesis, random walk, unbalanced active distribution system.**

Nomenclature

A. General Abbreviations

DER	Distributed energy resource.
D	Discriminator neural network.
G	Generator neural network.
KDE	Kernel density estimation.
LSTM	Long short-term memory.
MIQP	Mixed integer quadratic programming.
NET	Original distribution network.
PDF	Probability density function.
RW	Probability density function.
SM	Smart meter.
UG-GAN	Unbalanced graph generative adversarial network.

B. Parameters and Functions of Wasserstein GAN and UG-GAN

A	Adjacency matrix.
c	Clipping parameter.
$\mathbb{E}(\cdot)$	Expectation function.
f_θ	Kernel Sequential neutral network.
$g_{\theta'}(\cdot)$	Initialization parametric function.
m	Batch size.
n_{iter}	Number of discriminator iterations per generator iteration.
p_{real}	Possibility of real of input data x.
$\mathbb{P}_x$	Distribution of the real samples x.
$\mathbb{P}_z$	Distribution of the noise signal z.
Q	Scoring matrix.
T	Number of random walk step.
$V(\cdot)$	Value function.
v_i	Random walk vector of the i-th step.
x	Real data.
x_{fake}	Generated artificial data.
z	Noise signal data.
α	Learning rate.
θ_g	Learning parameter of G.
θ_d	Learning parameter of D.
θ_{g0}	Initial learning parameter of G.
θ_{d0}	Initial learning parameter of D.
$\mathcal{N}(\cdot)$	Multivariate Gaussian distribution.
$Cat(\cdot)$	Category function.
$\sigma(\cdot)$	Sigmoid function.

C. Parameters and Variables of Network Correction and Extension

E	A $N_e \times 2$ matrix indicating from and to node indexes of the i-th edge.
h_j	Kernel bandwidth for the j-th variable.

Manuscript received 18 December 2021; revised 18 April 2022 and 16 August 2022; accepted 1 October 2022. This work was supported in part by the National Science Foundation under Grant CMMI 1745451. The work of Rong Yan was supported by China Scholarship Council under Grant 201906320221. Paper no. TPWRS-01938-2021. *(Corresponding author: Zhaoyu Wang.)*

Rong Yan is with the College of Electrical Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China, also with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA, and also with the Power Dispatching and Control Center, China Southern Power Grid Co., Ltd., Guangzhou, Guangdong 510663, China (e-mail: yanrong052@zju.edu.cn).

Yuxuan Yuan and Zhaoyu Wang are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: yuanyx@iastate.edu; wzy@iastate.edu).

Guangchao Geng and Quanyuan Jiang are with the College of Electrical Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: ggc@zju.edu.cn; jqy@zju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2022.3212029.

Digital Object Identifier 10.1109/TPWRS.2022.3212029

$K_j(\cdot)$	j-th kernel function.
l_{ij}	Square of current.
v_j	Square of voltage.
r_{ij}	Resistance of line $i-j$.
x_{ij}	Reactance of line $i-j$.
n	Number of elements for a variable.
N_c	Number of user-defined library of conductor configurations.
t_s	Specific time slot.
$P_{E_{i1}E_{i2}}$	Active power of i-th transmission line connecting node E_{i1} and E_{i2}.
P_{Ij}, P_{Lk}	Active power load of j-th three-phase and k-th single-phase.
p_{Gj}, q_{Gj}	Active and reactive capacity of grid component at node j.
p_{Dj}, q_{Dj}	Active and reactive load at node j.
P_{ij}, Q_{ij}	Active and reactive power flow of line $i-j$.
$\cdot_{min}, \cdot_{max}$	User-defined thresholds of the given variable.
$\bar{\cdot}, \underline{\cdot}$	Upper and lower bound of the given variable.
$u^A_{Lk}, u^B_{Lk}, u^C_{Lk}$	Binary variables correspond to phase A, B, or C for k-th single-phase load connected to node M_{Lk}.
$u^A_{Ij}, u^B_{Ij}, u^C_{Ij}$	Binary variables correspond to phase A, B, or C for j-th three-phase load connected to node M_{Lk}.
$u^A_{n\zeta}, u^B_{n\zeta}, u^C_{n\zeta}$	Binary variables correspond to phase A, B, or C for node ζ.
$u^A_{ei}, u^B_{ei}, u^C_{ei}$	Binary variables correspond to phase A, B, or C for edge i.

D. Performance Evaluation of Generated Network

D_{avg}	Average node degree.
D_{max}	Maximum node degree.
D_{br}	Branching rate.
De_{max}	Maximum depth.
N_L	Number of single phase loads.
N_I	Number of three phase loads.
N_n	Number of distribution network nodes.
N_e	Number of distribution network edges.
$P_{L,avg}$	Average nodal active power of loads.
$P_{L,max}$	Maximum nodal active power of loads.
P_0, Q_0	Active and reactive power at the interface of transmission and active distribution network.
PF	Power factor.
ρ_{PC}	Assortativity coefficient.
Δ	Imbalance ratios of unbalanced distribution systems.

I. Introduction

POWER researchers seek to understand how real-world systems work and how real-world systems can work better. Therefore, knowledge of real-world systems, including topologies, locations and parameters of electrical components, and customer consumption behaviors, is essential to their works. In practice, most utilities are hesitant to share their systems with the public due to data privacy concerns. One common solution is to use IEEE test feeders modified on real distribution systems for model validation and demonstration. However, the main challenge is that the number of standard test feeders is very limited. Hence, synthetic test systems have been developed as alternatives to represent various real networks flexibly. Basically, synthetic networks should exhibit the critical topological and electrical characteristics of real-world networks with user's behaviors, but they are entirely fictitious, and users cannot extract any real-world network information from synthetic networks by reverse engineering.

Previous works mainly focus on generating synthetic transmission networks, which can be classified into two categories: statistics-based [1], [2], [3], [4], [5], [6], [7] and machine learning-based [8], [9] methods. The statistics-based methods performed extensive data analytics on a large amount of real-world power grid data to manually quantify the key properties, both topological and electrical, of network, such as node degree, load distribution, and parameters of grid components. Based on these properties, synthetic networks can be generated using graph theory and grid planning simulations. Specifically, reference [1] and [2] present the methods to get a set of statistical metrics by analyzing empirical probability density function of transmission network electrical parameters. These metrics are significantly important both in the network creation and validation stage. With these properties, reference [3] presents a systematic synthetic power grid creation method, which can be seen as a general solution for realizing this task. Latter researches based on statistics-based methods mostly focus on customizing a more realistic power grid for a specific study field, including testing the influence of geomagnetic disturbance [4], economic criteria [5], and communication and control network [6] on real-world gird. Instead of statistics-based methods, machine learning-based methods are also introduced in this field by predicting the connectivity of the grid directly according to the distribution of the training networks properties. In [8], [9], network imitating methods were proposed to generate grids with similar properties to the given networks. Both methods are based on the small-world assumption [7], which has been proved by most scholars in field of transmission systems, i.e., a type of system in which most buses are not directly connected, but the neighbor buses of any given bus are likely to be directly connected and most buses can be reached from other buses by a small number of buses. Recently, some works start rethinking whether small world is an accurate model for transmission grids [10] by a small group of researches, and attempt to design new techniques, e.g., methods based on system planning sensitivities [10], to produce a more realistic synthetic grid. It is worth noting that the network created by these methods is able to basically meet requirements of actual applications.

Compared to transmission network synthesis, research on active distribution network synthesis is still at a preliminary stage. Some studies [11], [12] have extended the transmission-level statistics-based methods to distribution grids by introducing several indices representing topological properties of distribution networks. However, distribution networks definitely no longer satisfy the small-world assumption, which impacts the performance of these methods. Moreover, the regional nature of the distribution systems is greatly ignored in these works. For example, urban and rural distribution networks have different properties in both topology properties and power flow distribution. Based on our observations of real-world data, the characteristics of distribution networks depend heavily on street layout, space availability, customer density, and even utilities' own preferences. Such observations indicate that each distribution network has a great deal of specificity. Consequently, some researchers have used local geographical and social statistic data, such as google maps and Census data, to simulate the system planning process for distribution network synthesis [13], [14], [15]. In fact, it is an alternative way since existing works cannot extract all key information for a specific distribution network. Although the best one among them [15] is able to create a realistic large-scale network, it still largely depends on the expert experience of planning and huge amount detailed local geographical, social statistic, and electrical data. Others try to develop representative synthetic test feeders directly from real systems using hierarchical clustering analysis manually. For example, in [16], 24 networks were presented from 575 real distribution feeders, which characterize distribution systems in different regions of the U.S. Apart from this, authors of [15] extend their works to synthetic combined transmission and distribution networks synthesis task [17] and do validate in an electrical manner, trying to build a more realistic synthetic grid with larger scale.

While previous works provide valuable insights, some challenges remain unanswered or only partly covered in this area and can be summarized as follows: (1) Existing statistics-based works [11], [12], [16] normally rely on a large amount of real-world data to extract statistical grid properties. Besides, other planning simulation-based works [13], [14], [15] also require a mass of detailed local geographical and social statistic data. Such a strategy not only poses a challenge for data acquisition and privacy but also raises concerns about the generalizability of the methods. When researchers generate synthetic grids for model development and validation, they need to first extend their datasets by collecting massive real-world data, which is very expensive. (2) Previous methods [11], [12], [16] ignore the significant diversity of distribution systems due to different geographic environments and grid infrastructures. For example, urban distribution systems show very different topological and electrical factors than rural systems.(3) For all existing works [11], [12], [13], [14], [15], [16], it is not well studied how to create realistic unbalanced active distribution systems, which is exactly one of the key features in practical distribution grids. (4) The previous works [11], [12], [13], [14], [15], [16] pay more attention to the grid connectivity generation, rather than the interaction between topology, loads, and electrical components. Besides, they do not provide time-series nodal load data reflecting the users' behavior, and it limits the scope of application scenarios.

To address these challenges, we propose a data-driven framework that uses limited real-world data to generate a comprehensive active distribution test feeder. Here, "comprehensiveness" means that it contains time-series nodal demands and standard distribution grid components with realistic parameters. To achieve this, first, an unbalanced graph generative adversarial network (UG-GAN) method is designed to produce synthetic node connectivity. Specifically, we formulate the network synthesis problem as learning the distribution of biased random walks[1] both over a single real-world network and across each phase of line segments. Also, we modify the standard GAN architecture to handle the discrete nature of the network data. When the UG-GAN is trained, synthetic node connectivity can be obtained by repeatedly generating random walks. Then, based on this synthetic topology, we utilize a non-parametric uncertainty quantification method known as kernel density estimation (KDE) to generate time-series load consumption data for each node. Finally, an optimization-based component placement model is proposed to determine the locations and parameters of various grid components. The goal of this optimization model is to consider the interactions between topology, loads, and electrical components in distribution systems. Unlike previous works that validate synthetic networks only in a statistical manner, our method is tested in a power system manner. More precisely, the generated test case is applied in three different power applications. Case studies demonstrate that our synthetic active distribution system has similar electrical properties and significantly different external characteristics to the input network, which respects the data autonomy of the data owner.

By using the proposed method, researchers and engineers can mimic one particular real-world network and generate a set of comprehensive testing cases with similar proprieties. As a result, data providers will no longer have any concerns about making desensitized data publicly available in response to requests from industry and academia. In other words, data providers will be more willing to share synthetic systems generated using our methods rather than sharing their real-world systems directly. Also, although this work is fine-tuned on our dataset to optimize the values of the model hyperparameters, the methodology is general and can be applied to any other radial distribution systems for system synthesis after retraining/fine-tuning to capture the unseen distribution of random walks. This is true for any data-driven solution. Furthermore, our model has good scalability. Specifically, the proposed method operates on random walks and only considers the non-zero entries of the adjacency matrix instead of generating the entire adjacency matrix, which requires computation and memory as a quadratic function of the number of nodes. Such a strategy efficiently exploits the sparsity of real-world active distribution systems to enhance scalability. Meanwhile, given that system synthesis is a purely offline analysis, the computation burden of the proposed UG-GAN does not directly impact the performance of our method.

In summary, the innovative contributions of this paper can be summarized as follows:

- The proposed model follows an adversarial generative framework that allows the use of limited real-world data (at least all key information of one real distribution network) to capture the specificity of individual three-phase unbalanced active distribution systems while maintaining confidential information.

[1] Biased random walk is a randomly sampled path that consists of a succession of random steps on a given graph. Unlike in a pure random walk, the probabilities of the potential new states are unequal due to the topology of the given graph.

- The proposed method can generate a comprehensive distribution test case that contains three-phase unbalanced topology, more detailed time-series nodal load data, and more types of standard grid components in order for broader application scenarios.
- Topological and electrical indices, together with three power applications, are introduced to verify that the generated active distribution systems are realistic.

II. UG-GAN Based Unbalanced Distribution Network Synthesis

In this section, a UG-GAN is proposed to generate unbalanced distribution networks by using a single network. To help the reader understand our model, we first review Wasserstein GAN, including basic idea, formulation, and training process, then describe the details of our UG-GAN.

A. Wasserstein Generative Adversarial Network

Wasserstein Generative Adversarial Network (Wasserstein GAN) is a novel GAN architecture [18] that improves the training stability and provides a loss function to describe the quality of the generated samples [19]. It is with the ability to learn the underlying distribution $\mathbb{P}_x$ of the real samples x, by finding out a mapping relationship from a known sampled distribution $\mathbb{P}_z$ (such as Gaussian distribution) to an artificial sample that follows $\mathbb{P}_x$. This function can be realized by two deep neural networks: a generator (G) and a discriminator (D). The interaction between these two networks is formulated as a game-theoretic two-player nested min-max optimization $V(G, D)$. For concreteness, they are described as follows:

1) Generator Neural Network (G): G defines an end-to-end neural network trained to transform a noise signal z to the generated artificial data x_{fake}:

$$x_{fake} = G(z; \theta_g) \tag{1}$$

where θ_g denotes the learning parameter of G. z is the noise signal with a known probability density distribution. In this work, we choose the noise with multivariate Gaussian distribution, shown as:

$$z = \mathcal{N}(0, z_\sigma) \sim \mathbb{P}_z \tag{2}$$

General speaking, any machine learning model (like artificial neural network, convolutional neural network, long short-term memory or ensemble model) can be embedded into G, according to the specific requirements of different tasks, so that the generated artificial data satisfies the distribution of real data $\mathbb{P}_x$.

2) Discriminator Neural Network (D): D is trained to maximize the probability of assigning the correct labels to both real examples and artificially generated samples from G. It outputs a single scalar p_{real} ranging from 0 to 1, representing the possibility that the input data x is from the real dataset rather than generated artificially by G. The network with learning parameter θ_d is listed as:

$$p_{real} = D(x; \theta_d) \tag{3}$$

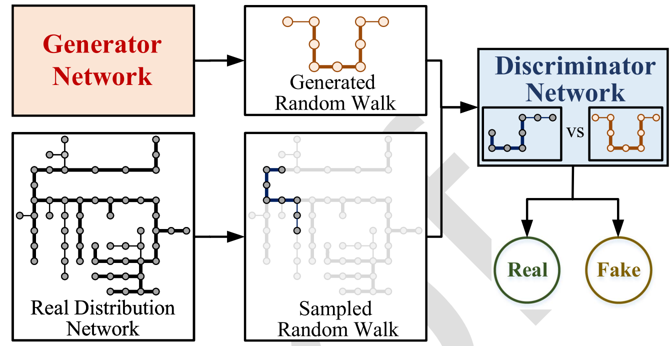


Fig. 1. Proposed UG-GAN architecture.

3) Value Function $V(G, D)$ and its Training Process: As mentioned above, G can be regarded as a model to learn a mapping relationship $G(z; \theta_g)$ from noise with known distribution to real data space. Thus, the training object is obviously to make the generated artificial data as realistic as the real ones from the perspective of D, by maximizing the expectation of generated artificial data $\mathbb{E}_z[D(G(Z))]$. Meanwhile, $D(x; \theta_d)$ is defined as another neural network to distinguish real data from artificial ones, with an objection maximizing the expectation difference between real data $\mathbb{E}_x[D(x)]$ and generated data $\mathbb{E}_z[D(G(Z))]$. Therefore, a suitable value function $V(G, D)$ for these two interconnected networks is the key idea of GAN, by modeling as a game-theoretic two-player minimax optimization problem. Noted that this value function is specially designed in Wasserstein GAN to improve the stability of the training process on the basis of traditional GAN, shown as:

$$\min_G \max_D V(G, D) = \mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z))] \tag{4}$$

Two networks are trained simultaneously via an adversarial process using the above value function, until reaching a unique global optimum. More details can be found in [18].

B. UG-GAN for Unbalanced Network Synthesis

In power systems, despite novel generative models have great success in dealing with real-valued data, such as wind and outage scenario generation [20], [21], adapting generative models to handle discrete network data is still an open problem. Therefore, in this paper, we propose a new algorithm, UG-GAN, to address the needs of our task. The main idea is illustrated in Fig. 1. Basically, the proposed model captures graphical features of a network by learning the distribution of biased random walks over the network. As demonstrated concretely in [22], random walk is a stochastic sampled path that consists of a succession of random steps on a given network. A distribution grid can be decomposed into a set of random walks that contain both local and global graphical features. Generally speaking, similar networks share similar distribution of sampled random walks, as long as the sampled random walks are sufficient. Following this theory, random walk sampling is employed to convert network data to sequential data.

1) Random Walk Sampling and its Encoding Scheme: To indicate the process of random walk sampling and encoding

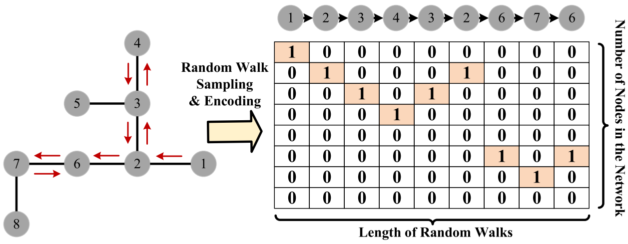


Fig. 2. An example of random walk sampling and its encoding scheme.

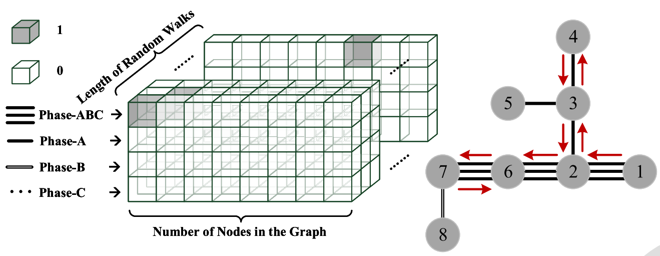


Fig. 3. Proposed two-dimensional one-hot encoding scheme for unbalanced distribution systems.

scheme, an 8-node radial network is illustrated as an example, as shown in Fig. 2. Here, we assume that each edge in this network has the same probability of being selected. For example, when the current position of the random walk is node 3, the probability of edge 3-2, 3-4, 3-5 being sampled at the next step is regarded as the same. It is worth noting that a single random walk does not necessarily include all system nodes. In this example, nodes 5 and 8 or edge 3-5 and 7-8 are not sampled. However, as the number of random walks increases, all nodes and edges will be sampled thousands times. It is clear that the distribution of these sampled random walks on graphs highly depends on the given graph topology. Also, as the number of nodes in the network increases, more random walks are required to cover the entire graph. As a result, the conversion from a grid topology to a set of random walks can be regarded as an equivalent transformation, and our task is no longer to learn the hidden features of grid, but the extract features from those random walk. Then, a one-hot encoding scheme is employed to further convert the random walk to the integer representation, as shown in the right part of Fig. 2.

Considering that unbalanced multi-phase distribution systems (e.g., with single and three-phase laterals) are prevalent in the U.S., we propose a new two-dimensional one-hot encoding scheme to embody the phase information of the input network. It extends one additional dimension for each random walk step based on one-hot encoding scheme, to determine the state including both node and phase information at the same time. In other words, we use a two-dimensional matrix to indicate the state of each random walk step. Note that for this matrix, only one element is equal to 1 to ensure consistency with grid physics. As shown in Fig. 3, the input of D and the output of G can be rearranged into a three-dimensional tensor for each random walk sampled from the original network. Specifically, the first two dimensions are represented by a two-dimensional matrix with N_n columns and four rows, denoting the phase information of each random walk step (phase A, B, C, and ABC respectively, moreover, if two-phase loads exist, it should be seven rows). The third dimension is the length of the random walk. For this example, as for the first layer of tensor in Fig. 3, only the elements in the first row and third column are equal to 1, which means the first node of this selected random walk is a three-phase node with index 1.

Apart from processing phase information, it is necessary to select line conductors and their configurations. To achieve this, a library of conductor types and configurations is used, which can be easily found in utility guidance for distribution systems under specific voltage levels [23], [24]. We have embedded this selection solution into the unbalanced topology synthesis process as a unified problem. In doing so, an additional dimension is added on the basis of the two-dimensional one-hot encoding scheme for the selection of line conductors with their configurations. Specifically, the input of the discriminator and the output of the generator in the proposed UG-GAN model can be rearranged into a four-dimensional tensor for each random walk sampled from the original network or generated by the generator network. The first two dimensions are represented by a two-dimensional matrix with N_n columns and N_c rows, where N_c is determined based on a user-defined library of conductor configurations. The third dimension is the length of the random walk. The fourth one denotes the four or seven possible types of phase information of each random walk step. In other words, it is extended to a three-dimensional one-hot encoding scheme, and each little square shown in Fig. 3 is split into several elements, representing all possible conductor types. In such case, conductor can be sampled, for each step of random walk, from the library of utility guidance using the same encoding scheme aforementioned, apart from determining a specific phase and network connectivity. Similar approaches can be further employed for other in-series grid component placement, which is seen as a special conductor type, like circuit break, regulator, and etc.

2) Structure of Generator Neural Network in UG-GAN: Given an input distribution network, defined by a binary adjacency matrix, $A \in \{0,1\}^{N_n \times N_n}$, we first sample a large number of random walks $RW := \{v_1, v_2, \ldots, v_T\}$ of length T from A. Then, these random walks are used as the training set of G, which can be formulated as follows:

$$(h_t, C_t, p_t) = f_\theta(h_{t-1}, C_{t-1}, v_{t-1}) \quad (5a)$$

$$v_t \sim Cat(\sigma(p_t)) \quad (5b)$$

$$(h_0, C_0) = g_{\theta'}(z), \quad v_0 = 0 \quad (5c)$$

where $\sigma(\cdot)$ is the sigmoid function, $Cat(\cdot)$ is a category function, and $g_{\theta'}(z)$ denotes a parametric function from the noise signal generated by the multivariate Gaussian distribution to initialize a sequential neutral network f_θ. In this work, a modified long short-term memory (LSTM) is utilized to represent f_θ. As shown in Fig. 4, for each time step t, LSTM cell outputs two values: current state vector h_t and C_t, and discrete possibility vector p_t for all possible nodes to be sampled at the next time step $t+1$. Since sampling from a categorical distribution is the non-differentiable operation that impedes backpropagation, we have

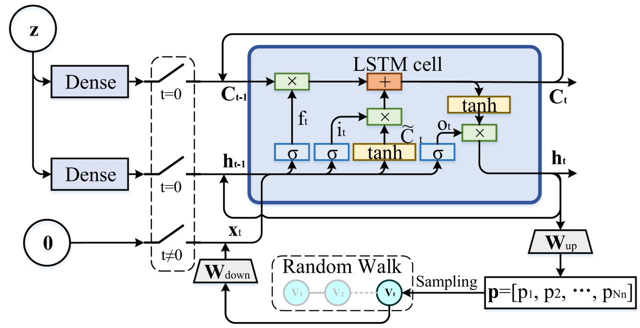


Fig. 4. LSTM-based generator architecture of UG-GAN.

applied the Gumbel-Max trick to solve this problem [25]. After relaxation, the exact node v_t of random walk can be sampled according to p_t using (5b).

3) Structure of Discriminator Neural Network in UG-GAN: D is based on the standard LSTM architecture to distinguish sequential random walks generated by G from the ones sampled from the real distribution network. Further, an input data preprocessing and an output activation layer are added to D. More precisely, at each time step, the random walk vector v_t encoded in a two-dimensional one-hot format is reshaped before fed into LSTM as input. The output of the discriminator is a scalar indicating the probability that the input random walk is real.

4) Training Algorithm of UG-GAN: In this subsection, we present the training algorithm of UG-GAN giving the reader a clear picture of the training process. Frankly speaking, the training process of UG-GAN follows the line of Wasserstein GAN [18] with minor modifications, as it prevents mode collapse and leads to more stable training. As shown in Algorithm 1, it requires the original distribution network with several parameters as input, and outputs the final parameters of G and D.

After the training process, G can implicitly represent the underlying distribution of biased random walks over the real-world network and D cannot distinguish the true random walks from the artificial random walks. The biased second-order random walk sampling strategy described in [26] is utilized in G. Based on the random walks generated by G, a scoring matrix Q is constructed, not only measuring the possibility of connectivity for each node, but also providing phase information and conductor configurations if connected.

Algorithm 1: UG-GAN Training Algorithm.

Require:
α, the learning rate.
c, the clipping parameter.
m, the batch size.
n_d, the number of iterations of the discriminator per generator iteration.
θ_{d0}, initial discriminator's parameters.
θ_{g0}, initial generator's parameters.
NET, original distribution network.

Output:
θ_d, parameters of discriminator.
θ_g, parameters of generator.

```
Sample a huge amount of random walks RW from the input distribution network NET, and encoding them as the real input dataset x.
while not converged do
  for n_iter = 0, ..., n_d do
    Sample {x^(i)}_{i=1}^m ~ P_x a batch from the real data.
    Sample {z^(i)}_{i=1}^m ~ P_z a batch of prior samples.
    G_{θ_d} ← ∇_{θ_d}[1/m Σ_{i=1}^m D(x^(i); θ_d) − 1/m Σ_{i=1}^m D(G(z^(i); θ_g); θ_d)]
    θ_d ← θ_d + α · RMSProp(θ_d, G_{θ_d})
    θ_d ← clip(θ_d, −c, c)
  end for
  Sample {z^(i)}_{i=1}^m ~ P_z a batch of prior samples.
  G_{θ_g} ← −∇_{θ_g} 1/m Σ_{i=1}^m D(G(z^(i); θ_g); θ_d)
  θ_g ← θ_g − α · RMSProp(θ_g, G_{θ_g})
end while
```

III. Active Distribution Network Correction, Extension and Evaluation

When the graphical features of the real-world network are captured by the UG-GAN, an active distribution network correction and extension framework is developed to provide a comprehensive distribution test case, including realistic nodal load data and standard grid components with detailed parameters.

A. Time-Series Load Data Synthesis

The basic idea of load data synthesis is to estimate the probability density of multiple load behaviors and then sample them accordingly. However, considering the highly complex load uncertainty, it is difficult to do utilizing traditional parametric density estimation methods with Gaussian, beta, and GMM distribution model assumptions. This is because these methods rely on model assumptions that may introduce significant modeling bias in uncertainty quantification.

To address this challenge, a non-parametric method, known as kernel density estimation (KDE), is employed to estimate the probability density function (PDF) of different load behaviors,[2] and generate the time-series load data for each primary nodes by sampling the estimated PDFs. For concreteness, the proposed algorithm is summarized as three steps. The first step is to collect the time-series load data of all types of users. Then, these load data are classified using an unsupervised clustering algorithm [27] to reduce the uncertainty of load behaviors. For each type of customer, the Davies-Bouldin validation index (DBI) is utilized [28] to determine the optimal number of clusters. The relational behind DBI is to quantify the ratio of within-cluster and between-cluster similarities. The second step is to estimate load PDF of each cluster. Let X is a matrix of d variables drawn from the load distribution of a cluster with an unknown density

[2]Electricity customers can be roughly divided into three main types with completely different consumption behaviors: residential, commercial and industrial loads.

f, which is formulated by:

$$f(x_1, x_2, \ldots, x_d) = \frac{1}{n \cdot \prod_{j=1}^{d} h_j} \sum_{i=1}^{n} \prod_{j=1}^{d} K_j \left(\frac{x_j - X_{ij}}{h_j} \right) \tag{6}$$

where n donates the number of elements for a variable, and h_j is the kernel bandwidth for the j-th variable. $K_j(\cdot)$ is the j-th kernel function. The Gaussian kernel function is adopted here. In this application, load data is considered as a time varying variable, thus for a specific time slot t_s, the conditional density function of per unit load can be expressed as:

$$f(P_L|t = t_s) = \frac{f(P_L, t = t_s)}{f(t = t_s)} \tag{7}$$

The final step is to generate synthetic load data by sampling from each PDF. To further protect data privacy, we only provide the nodal load data of generated primary network rather than each end user.

For distribution systems with distributed energy resources (DERs) and renewable distributed generators (DGs), the algorithm aforementioned can also be employed with minor modifications. It is worth noting that most DGs and part of DERs are invested by power consumers, and they are normally installed behind-the-meter. In this case, when the utilities only have the smart meter data, the proposed method can be first utilized to generate net demand and then applied the data-driven disaggregation methods, such as our previous work [29] desgined for residential rooftop solar photovoltaics (PVs), to obtain power generation data and native demand. If the utilities install devices to monitor the solar generation, our method can be applied to generate two time series: one for native demand and one for renewable generation. The main difference is that the generator output data is clustered according to the scenarios, like weather events (e.g., high wind day, sunny day), generator types, and other influential factors.

B. Load Assignment and Topology Correction

By using on the Q matrix generated by UG-GAN, one simple solution determining the topology, phase information and conductor configurations of the synthetic network is to choose the edges and their corresponding line conductors with the highest probability. However, such a solution does not take into account the strong coupling relationship between topology and load distribution, which leads to significant differences between generated networks and actual grids. For example, in practical systems, utilities prefer to connect industrial customers with an individual three-phase node (without other residential customers) to ensure the reliability of the power supply. Besides, there may also be restrictions in the selection of upstream and downstream[3] line conductors. Therefore, an optimization-based joint framework of load assignment and topology correction is proposed, in order to assign all loads to the generated system while performing topology corrections. Specifically, this joint framework is cast as a Mixed Integer Quadratic Programming (MIQP) problem. Among them, 12 binary variables are defined to represent the connectivity between loads (including N_L single-phase and N_I three-phase loads) and the generated network with N_n nodes and N_e edges:

$$\begin{aligned} &u_{Lk}^A, u_{Lk}^B, u_{Lk}^C, u_{Ij}^A, u_{Ij}^B, u_{Ij}^C, u_{n\zeta}^A, \\ &u_{n\zeta}^B, u_{n\zeta}^C, u_{ei}^A, u_{ei}^B, u_{ei}^C \in \{0, 1\} \end{aligned} \tag{8}$$

The first three variables correspond to individual phase for k-th single-phase load connected to node M_{Lk}, where $k = 1, 2, 3, \ldots, N_L$. The fourth to sixth variables indicate individual phase for j-th three-phase load connected to node M_{Ij}, where $j = 1, 2, 3, \ldots, N_I$. The last six denote individual phase for node ζ and edge i, respectively.

First, optimization objective is formulated as follows to determine a final network according to matrix Q.

$$Obj = \sum_{i=1}^{N_e} ((Q_{i,1} - u_{ei}^A)^2 + (Q_{i,2} - u_{ei}^B)^2 + (Q_{i,3} - u_{ei}^C)^2) \tag{9}$$

Second, several constraints are added to ensure consistency with grid physics. We will describe them one-by-one. For the k-th single-phase load, it can merely be assigned to a specific phase of the network. Thus, the constraints corresponding to the binary variables of each phase can be written as:

$$u_{Lk}^A + u_{Lk}^B + u_{Lk}^C = 1 \tag{10}$$

In a similar manner, for the j-th three-phase load which is connected to the ζ-th node, the constraints of phase-A binary variables can be written as:

$$u_{Ij}^A = 1, \quad u_{n\zeta}^A = 1 \quad j \in \zeta \tag{11}$$

For all customers connected to node M_{Lk} ($k \in \zeta$), the binary variables associated with the load and node satisfy boolean logical relationship "or". We use phase-A as an example to explain this: u_{nk}^A will be 1 when a single-phase load connects to phase-A of this node or a three-phase load connects to this node, otherwise it will be 0. In this work, we convert this boolean operation to a set of constraints as follows:

$$u_{n\zeta}^A \geq u_{Lk}^A, \quad u_{n\zeta}^A \leq \sum_k u_{Lk}^A, \quad k \in \zeta \tag{12}$$

Considering that the vast majority of distribution networks in normal operation are tree-like structures [12], the upstream and downstream edges and nodes in the generated topology should meet several rules. Obviously, when the upstream edge is a three-phase branch, the downstream one can be either a single or three-phase branch. In contrast, the downstream one can merely be a single-phase branch when the upstream edge is a single-phase branch. Meanwhile, the phase information of the downstream node should be aligned with that of upstream edges. These two rules are formulated as a set of constraints described in (13).

$$u_{nE_{i1}}^A \geq u_{nE_{i2}}^A, \quad u_{ei}^A = u_{nE_{i2}}^A \tag{13}$$

[3] The upstream and downstream relationships of the nodes and edges are used to define the power flow properties of the distribution network. For example, for a conductor with the power flow from node A to B, we name A as the upstream node and B as the downstream one. The definition of upstream and downstream edges are similar. In other words, we choose to use the concept of upstream and downstream nodes and edges to indicate the electrical properties as well as the topological characteristics.

where E donotes a $N_e \times 2$ matrix. In the i-th row, first and second column elements, E_{i1} and E_{i2} ($E_{i1} < E_{i2}$), are the from and to node indexes of the i-th edge, $i = 1, 2, 3, \ldots, N_e$. Further, a constraint is added to the model for avoiding overloads in the generated synthetic network:

$$\underline{P_{E_{i1}E_{i2}}} <= P_{E_{i1}E_{i2}} <= \overline{P_{E_{i1}E_{i2}}} \tag{14}$$

where, $P_{E_{i1}}E_{i2}$ indicates the active power of i-th transmission line connecting node E_{i1} and E_{i2}. $\overline{P_{E_{i1}}E_{i2}}$ and $\underline{P_{E_{i1}E_{i2}}}$ are the upper and lower bound of the active power of the certain line. Finally, the following equations are added as constraints on the model in order to prevent unreasonable three-phase imbalance ratios in the synthetic network:

$$P^A = \sum_j \frac{1}{3} u_{Ij}^A P_{Ij} + \sum_k u_{Lk}^A P_{Lk} \tag{15}$$

$$\Delta_{\min} \leq \Delta = \frac{\max\{|P^A - P^B|, |P^A - P^C|, |P^B - P^C|\}}{P^A + P^B + P^C} \leq \Delta_{\max} \tag{16}$$

where P_{Ij} and P_{Lk} are the j-th three-phase and k-th single-phase active power load. Δ_{min} and Δ_{max} are the user-defined thresholds, near the imbalance ratios of original real-world unbalanced distribution systems.

C. Extension of Network With Grid Components

The proposed UG-GAN with the network correction process can generate a synthetic active distribution network with the related nodal consumption data. However, without standard grid components, the synthetic distribution system cannot be treated as a comprehensive test case. Thus, in this work, by imitating the real planning process, a Mixed Integer Second-order Cone Programming (MISCP) problem is formulated to place several grid components, including capacitor banks and distributed energy resources (DER), on the basis of the synthetic network. The objective function is written to minimize the power losses as follows:

$$\min \sum_{(i,j)\in E} r_{ij} l_{ij} \tag{17}$$

where r_{ij} denotes the resistance of line $i - j$, $l_{ij} = |\boldsymbol{I}_{ij}|^2$, i.e. the square of current, and $\forall(i,j) \in E$. Obviously, reducing network losses is not the only factor to be considered in grid component planning. Some components are directly invested by customers with the goal of local economic optimization. Therefore, the objective function described above can be modified according to the actual needs of the generated synthetic networks. One point to note is that the modified function must still be a linear function of l_{ij} and u_j to ensure the solvability of the formulated MISCP optimization problem.

Further, this optimization problem should be subject to multiple constraints to force the installed components to be realistic. In general, the constraints of this optimization problem can be divided into two parts. The first part shown in (18) restricts the location and capacity of each grid component. Among them, the first two inequality constraints restrict the active and reactive power injections of each grid component to be equipped. The third one describes the overall limits of active power, determining the possibility of power flow reversal. The last constraint refers to the limitation of the component number.

$$\begin{cases} u_j \underline{p_{Gj}} \leq p_{Gj} \leq u_j \overline{p_{Gj}}, \quad j \neq 0 \\ u_j \underline{q_{Gj}} \leq q_{Gj} \leq u_j \overline{q_{Gj}}, \quad j \neq 0 \\ \sum\limits_{j\in N_n, j\neq 0} p_{Gj} \leq \epsilon_p \sum\limits_{j\in N_n} p_{Dj} \\ \sum\limits_{j\in N_n, j\neq 0} u_j \leq N_G \end{cases} \tag{18}$$

where u_j is a binary variable indicating whether the grid component with active capacity p_{Gj} and reactive capacity q_{Gj} is installed at node j. p_{Dj} is the active load at node j. $\overline{\bullet}$ and $\underline{\bullet}$ are the upper and lower bound of the variable.

The second part is the power flow constraints of the synthetic network. Considering that classic power flow constraints are non-linear equations, the overall optimization problem can only be formulated as a mixed integer non-linear programming problem, which is hard to solve. To alleviate such difficulty, a relaxed branch flow model [30] is employed in this subsection, which is thus modeled so as a set of second-order cone constraints as follows:

$$\begin{cases} p_j = \sum_{k:j\to k} P_{jk} - \sum_{i:i\to j}(P_{ij} - r_{ij} l_{ij}) + g_j v_j \\ q_j = \sum_{k:j\to k} Q_{jk} - \sum_{i:i\to j}(Q_{ij} - x_{ij} l_{ij}) + b_j v_j \\ v_j = v_i - 2(r_{ij}P_{ij} + x_{ij}Q_{ij}) + (r_{ij}^2 + x_{ij}^2) l_{ij} \\ \left\| \begin{matrix} 2P_{ij} \\ 2Q_{ij} \\ l_{ij} - v_i \end{matrix} \right\|_2 \leq l_{ij} + v_i \\ \underline{V_j^2} \leq v_j \leq \overline{V_j^2} \\ \underline{I_{ij}^2} \leq l_{ij} \leq \overline{I_{ij}^2} \\ p_j = p_{Gj} - p_{Dj} \\ q_j = q_{Gj} - q_{Dj} \end{cases} \tag{19}$$

where $v_j = |\boldsymbol{V}_j|^2$, P_{ij} and Q_{ij} are the active and reactive power flow of line $i - j$, x_{ij} is the reactance of line $i - j$.

Overall, various standard grid components, e.g., capacitor banks and DERs, are placed in this generated synthetic network using the proposed network extension method, changing or even reversing the distribution of synthetic network power flow. It enables the generated synthetic network is similar to a realistic active distribution network. It should be noted that the proposed network extension method cannot be integrated with our UG-GAN algorithm because the goal is to mimic a specific network rather than replicate the original network.

D. Performance Evaluation

In order to evaluate the performance of the proposed method, topological and electrical indices are defined as follows. Moreover, several power applications are introduced in this subsection to further demonstrate that our synthetic networks are useful for power researchers and utility engineers, replacing the unavailable real-world data.

1) Topological and Electrical Indices: Based on previous work [12], several statistical and electrical based metrics are utilized in both graph and power aspects to prove that our model reproduces the most known properties inherent to real-world networks, which are listed below:

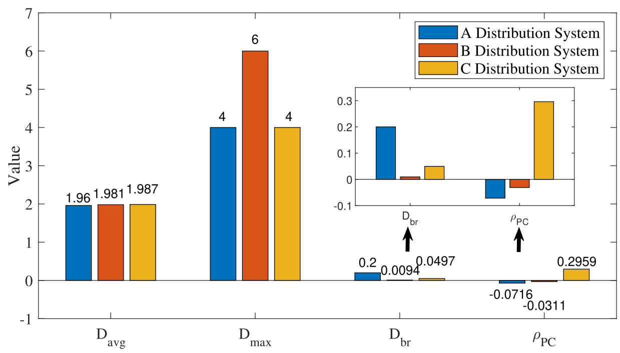


Fig. 5. Four statistical indices for the three distribution systems.

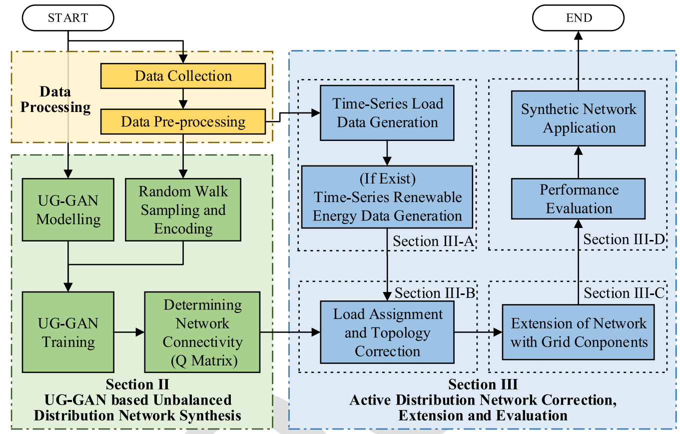


Fig. 6. Flow chart of the proposed method.

- N_n, N_e: The number of nodes and edges of synthetic active distribution network, which reflect the scale of the network.
- D_{avg}, D_{max}, D_{br}, ρ_{PC}: These four node degree-based indices are average node degree, maximum node degree, branching rate and assortativity coefficient, respectively. Among them, node degree represents the number of edges that are incident to a certain node, branching rate denotes the percentage of the number of nodes with degree greater than three, and assortativity coefficient is examined in terms of node degrees using the Pearson Correlation coefficient. These indices reflect the local graph properties of the active distribution systems. For example, urban or higher voltage level networks normally tend to branch out more compared to rural or lower voltage level ones.
- De_{max}: Maximum depth. It can be used to roughly describe the strength of the voltage drop in radial distribution systems.
- $P_{L,avg}$, $P_{L,max}$: Average and maximum nodal active power of loads, which reflect the baseline load level of the generated network.
- Δ: Three-phase unbalanced ratio defined in (16). This index reveals the unbalanced degree of the network.
- P_0, Q_0: Active and reactive power at the interface of transmission and active distribution network.
- PF: Power factor of the generated system.

Meanwhile, to prove that our model is not to simply replicate the original network, the ratio of overlapping edges (R_{oe}) between the real system and our synthetic system.

2) *Application Verification:* To further demonstrate that our generated active distribution network is realistic and useful, we review a question, that is, how to truly define whether the generated network is successful or not. It is indeed a more challenging problem, even compared to the network synthesis task. Most of the previous works only rely on statistical indices, obtained from a large amount of real-world data [1], [2], [3], [4], [5], [6], [7], [11], [12], [16]. However, as we mentioned before, topology properties are quite different for various distribution networks. This can also be confirmed using real data, as shown in Fig. 5. This figure shows four different indices (i.e., D_{avg}, D_{max}, D_{br}, ρ_{PC}) for the three distribution systems in the same region. It is clear that the statistical indices of the three systems are quite different, especially for D_{br} and ρ_{PC}. Thus, synthetic distribution system should be generated by a single network. Moreover, even if the statistical indices of synthetic networks are similar to those of real networks, it is difficult to guarantee that these networks can be used as alternatives for representing real networks. In our view, synthetic networks should be validated in a power system manner. In other words, the synthetic networks generated by our method should achieve similar results as the real network in various power applications. Hence, we have tested three common applications: power flow analysis, DERs placement, and transmission and distribution power flow co-analysis. Among them, power flow analysis is performed to verify that the synthetic system satisfies static stability limits, including voltage and line power flow limits. Besides, DERs placement and transmission and distribution power flow co-analysis are carried out to demonstrate that the co-operation of transmission system and active distribution network with partial reverse power flow is of no abnormality.

IV. Active Distribution System Synthesis Framework

In this section, we summarize the the proposed framework as a flowchart shown in Fig. 6, so as to present a clear view of the methodology. It can be observed that the whole process is divided into three parts: data processing stage, UG-GAN based network synthesis stage, and network correction, extension and evaluation stage.

In the first stage, data processing stage, the data needs to be collected and pre-processed in order to prepare for the latter two phases. Priority to listing the detailed data requirements, we should emphasize the purpose and the high-value use case of this paper again. When system operators need to share their networks and data with researchers or the third agents but have user privacy concerns, they can perform the proposed method to obtain the corresponding synthetic networks for different networks separately. Considering that different distribution network may share different properties, all we need to generate a synthetic network is all key information of a single real-world network. The detailed information to be collected is as follows:

1) Detailed three-phase unbalanced distribution network topology information with its parameter, including

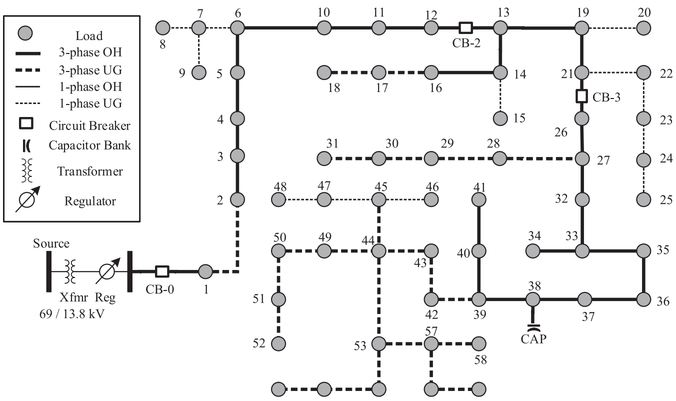


Fig. 7. Diagram of the original unbalanced distribution network.

network connectivity, phase information (phase A, B, C, AB, AC, BC, or ABC), and conductor parameters.

2) Time-series load data of different types of customers, including residential, commercial and industrial loads.
3) Grid components information, including in-series grid components (e.g., transformer, circuit break and regulator) and in-parallel grid components (e.g., capacity banks and DERs) (if exists).
4) Renewable distributed generators (DGs) information of the given distribution network (if exists).
5) Other optional components.

It is worth noting that data can be easily collected from utilities and their widely equipped smart meters. Then, data should be pre-processing (e.g., data cleaning), according to the requirement of the latter stages.

In terms of the second stage, a UG-GAN is proposed to synthesize unbalanced distribution networks by using the collected network topology and in-series grid components information. Specifically, a large set of random walks are sampled and encoded using the proposed method described in Section II-B1), making sufficient data preparation for UG-GAN training. At the same time, the generator and discriminator neural networks of UG-GAN are modeled respectively, using the proposed method introduced in Section II-B2) and 3). After that, the UG-GAN can be trained, and we can get the scoring matrix Q accordingly, determining the synthetic network connectivity.

As for the last stage, active distribution network correction, extension and evaluation process is developed to provide a final comprehensive distribution test case. This stage can be separated to four sub-parts. First, time-series load data and renewable distributed generators data is estimated using the historical dataset and KDE method, introduced in Section III-A. It is worth noting that these users or distributed generators are mostly installed behind-the-meter, so as we can get the detailed time-series historical data at the first stage. Second, the estimated load data is required to be assigned to the corrected network topology obtained from Section II, using the MIQP problem formulated in Section III-B. Besides, in order to extend the network to a more realistic and comprehensive test case, grid components need to be placed in the network aforementioned, with the approach described in Section III-C. So far, a distribution network has been obtained, with the similar electrical properties as the original network without privacy concerns. Finally, as introduced in Section III-D, the performance of synthetic network is evaluated using topological and electrical indices, together with power applications.

V. Case Study

This section explores the effectiveness of our proposed data-driven unbalanced network synthesis method by means of a case study. As detailed below, a 60-bus synthetic three-phase unbalanced distribution network is generated. Our simulation is mostly implemented in TensorFlow, an open-source machine learning platform, while optimization part in MATLAB with Yalmip and Cplex package. All cases are tested on a standard computer with Intel Core i7-8850H 2.6 GHz CPU, 16 GB RAM.

A. Data Requirements

In this section, the input system is a real-world distribution network data obtained from a Midwest U.S. utility [31], shown as Fig. 7. It is supplied by a 69 kV substation with 60 primary nodes and various grid components such as capacitor banks and line switches. Detailed information required to be collected in this case study is listed as follows.

1) The topological information of the 60-bus distribution network aforementioned.
2) Time-series load data of all types customers.
3) 5 types of grid components, including, transformer, circuit break, regulator, capacity banks and DERs.

Besides, in the UG-GAN training process, all random walks are sampled from this specific distribution network, with the same number as the ones generated from G network. In this case study, we sampled 128 random walks per iteration for discrimination and training in UG-GAN.

B. Distribution Network Synthesis Results

In this subsection, the detailed synthesis process is illustrated, and selected statistical and electrical-based indices are compared with the real-world input network, in order to verify the proposed method.

1) Visualization of Topology Synthesis Process: In the first few iterations of UG-GAN training, G and D of UG-GAN are still in a preliminary state with the initial parameters, as shown in Fig. 8(a). As a result, the generated network has many drawbacks, like isolated nodes, circle topology and etc. Then, in the early stage of the UG-GAN training process, when G performs poorly (the generated network is quite different from the real one), D can reject the generated random walks with a high degree of confidence. Therefore, in this stage, the discriminator loss drops dramatically to a small value, as shown in Fig. 9. After that, the two deep neural networks of UG-GAN are updated simultaneously via the adversarial process so that a more realistic topology can be generated, as shown in Fig. 8(b)–(g). When the training process is iterated 3,000 times, see Fig. 8(h), all topological properties of the generated distribution network are similar to those of the original network.

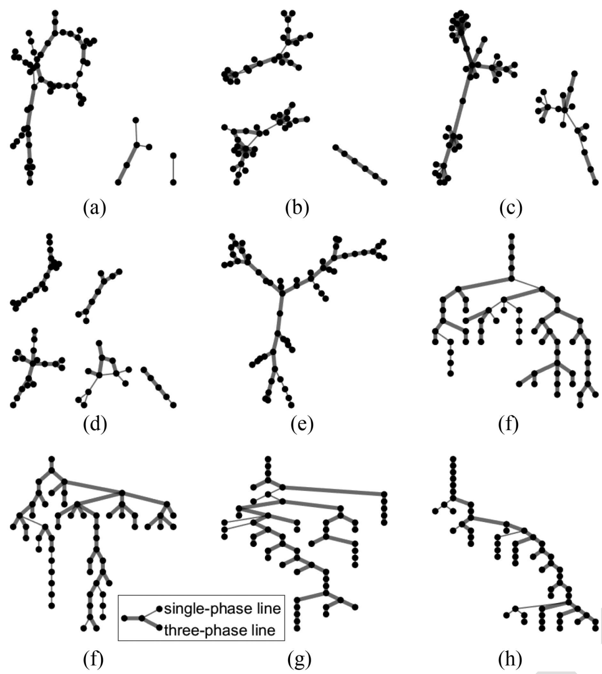


Fig. 8. Training process of UG-GAN.

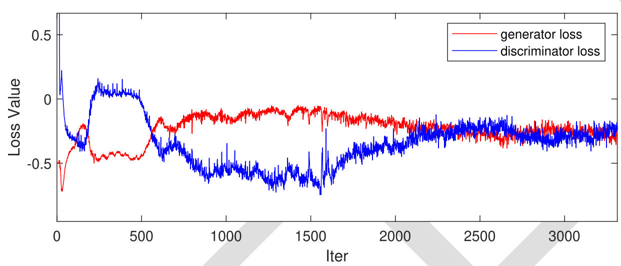


Fig. 9. UG-GAN Loss values of generator and discriminator.

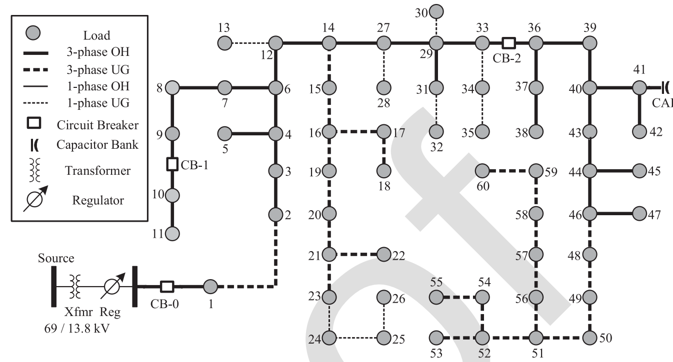


Fig. 10. Diagram of the synthetic unbalanced distribution network.

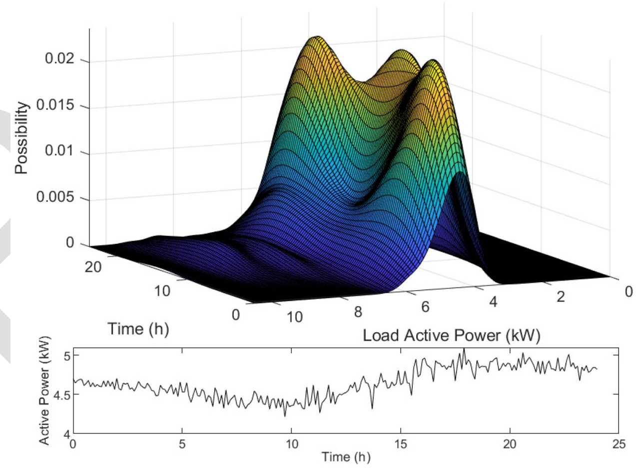


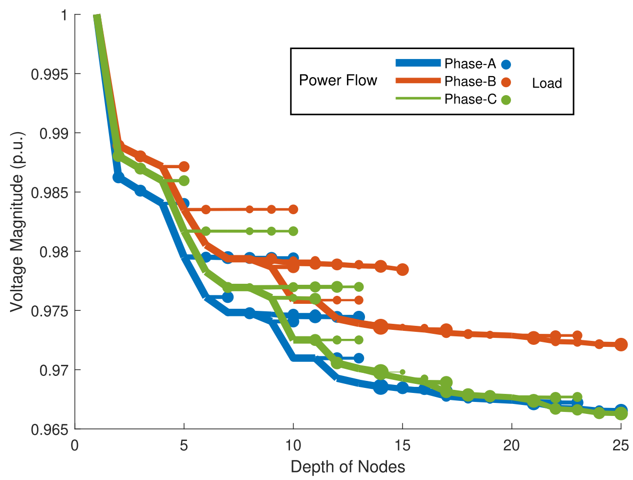
Fig. 11. Probability density diagrams of the residential customers.

Note that all edge-related information is determined at this stage by using UG-GAN, including distribution grid components (like circuit breaks) connected in series, cable type of each line, and etc.

2) Result of Load Data Synthesis Process: By using the proposed KDE-based method, the time-series data of 504 single-phase loads and 5 three-phase loads are generated and assigned to a certain phase on one of the 60 nodes in the generated network aforementioned. Fig. 11(a) and (b) illustrate the probability density diagram of a residential load and sampled time-series load data, respectively. To eliminate the possible customer's private information, the available customer power measurements are aggregated at the secondary transformer level by summing them at different times. Then, nodal loads are assigned to a certain phase of the generated network with minor topology correction using the formulated MIQP optimization problem to ensure the unbalanced degree within certain limits.

3) Synthetic Distribution System Description: The generated synthetic unbalanced distribution network consists of a 13.8 kV 60-node primary feeder that is supplied by a 69-kV substation. In this network, there are 57 branches in total, 48 of which are three-phase branches using 4 types of overhead lines and underground cables, and 9 of which are single-phase branches with 3 types of single-phase cables. The total length of the synthetic system is 3.34 miles. The three different types of unbalanced loads are assigned to 46 different nodes via secondary distribution transformers. Among them, an industrial three-phase load is connected to node #41, and residential or commercial loads are mixed together and connected to other nodes. Based on the results of our optimization-based component placement model, a capacitor bank is equipped near node #41 to provide reactive power support. Besides, 3 normally-closed circuit breaks are equipped in this network on lines 0-1, 9-10, and 33-36. The detailed structure of the generated network is illustrated in Fig. 10.

4) Indices Comparison of Generated Network: When the synthetic network is obtained, the aforementioned indices, indicating both topological and electrical properties, are used to compare the original and generated distribution networks, as shown in Table I. It can be clearly observed that all the representative statistical and electrical indices are similar. Meanwhile, the ratio of overlapping edges between two networks is about 0.5, preventing extracting real network confidential information by reverse engineering. Considering that the use of visualization can improve the interpretation of the results, we present the two networks directly, as shown in Fig. 7 (original network)



*The size of the solid circle represents the load of the node, and the thickness of the line indicates the size of the line power.

Fig. 12. Voltage and power flow of the synthetic distribution network.

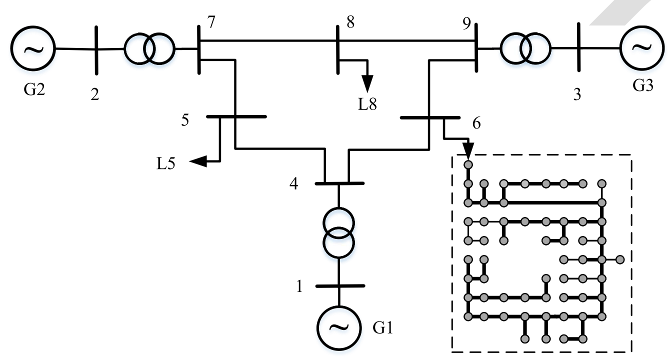


Fig. 13. Test system of transmission and distribution network co-simulation.

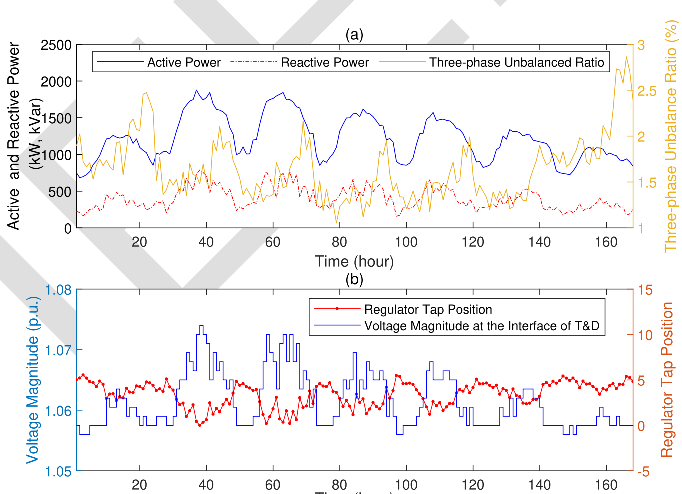


Fig. 14. Test result of transmission and distribution system co-analysis.

and Fig. 10 (synthetic network), so as to directly visualize the differences between the two networks.

To further demonstrate the effectiveness of our approach, we have conducted qualitative and numerical comparisons with the existing work. It is worth noting that the proposed method focuses on mimicking one particular network without any context data assumptions, e.g., local geographical and social statistic data, which poses a challenge for comparison with existing statistical-based methods. Also, the generated network on these methods are normally the three-phase balanced grid. Hence, to ensure a fair comparison among the existing grid synthesis method, we have compared the proposed method with a random tree algorithm that is the only method without involving any context data [32]. Specifically, by using this method, 50 different synthetic networks have been generated for investigation, as shown in Table II. Obviously, the topological indices of the synthetic networks generated by the previous method are far from the original network, especially for D_{br}, ρ_{PC} and De_{max}. Moreover, based on our observations, almost all randomly generated networks fail to satisfy the physical laws of the actual distribution system. For example, the upstream edge is a one-phase branch while the downstream one is a three-phase branch. Thus, such synthetic networks cannot be used to represent real-world systems in power system studies. As for other methods, including statistics-based method and simulated planning-based method, our proposed method is completely different in terms of purpose, algorithm, and input and output data. Thus, it is difficult to perform a fair quantitative comparison, since other

TABLE I
COMPARISON OF TOPOLOGICAL AND ELECTRICAL PROPERTIES BETWEEN THE GENERATED AND ORIGINAL DISTRIBUTION NETWORK

Topological Indices	Original Network	Generated Network
N_n	60	60
N_e	59	59
D_{avg}	1.9667	1.9667
D_{max}	4	4
D_{br}	0.0167	0.0167
ρ_{PC}	-0.0563	-0.0695
De_{max}	26	24
R_{oc}	/	0.508
Electrical Indices		
$P_{L,avg}$	25.56 kW	25.56 kW
$P_{L,max}$	1084.80 kW	1253.06 kW
PF	0.9834	0.9882
Δ	2.9%	2.5%
$P_0 + jQ_0$ (Phase-A)	614.04kW+j147.17kVar	585.89kW+j91.44kVar
$P_0 + jQ_0$ (Phase-B)	588.84kW+j98.32kVar	632.01kW+j99.74kVar
$P_0 + jQ_0$ (Phase-C)	642.04kW+j97.69kVar	626.87kW+j96.08kVar

TABLE II
COMPARISON OF INDICES OF DISTRIBUTION NETWORK BETWEEN DIFFERENT METHODS

Topological Indices	Original Network	Generated Network	
		Proposed Method	Random Method
N_n	60	60	60
N_e	59	59	59
D_{avg}	1.9667	1.9667	1.9667
D_{max}	4	4	4
D_{br}	0.0167	0.0167	0.0500
ρ_{PC}	-0.0563	-0.0695	-0.1394
De_{max}	26	24	16
R_{oc}	/	0.508	0.1186

* For the random method, we merely illustrate the indices results of the most similar synthetic network from all 50 randomly generated networks.

TABLE III
QUALITATIVE COMPARISON BETWEEN PROPOSED METHOD AND SIMILAR APPROACHES

	Proposed Method	Statistics-based Method	Simulated Planning-based Method
Representative References	/	[11], [12]	[13]–[15]
Purpose	1. Share real-world networks but with privacy concerns 2. Provide cases for research or algorithm test 3. Perform transmission and distribution system co-analysis	Provide a completely new synthetic distribution network for research or algorithm test	
Algorithm	UG-GAN	Quantify key statistical properties manually, and use graph theory and grid planning simulation	Use collected geographical and social data for grid planning simulation from scratch
Input Data	All key information of one given network	Thousands of distribution networks	Local geographical and social data, expert's experience, user type data, and etc.
Output Data	Synthetic network with similar features to original one, time-series nodal load data, and grid components data	A completely new synthetic distribution network which has nothing to do with any other existing real-world system	

methods cannot obtain a synthetic network using the same input data as our method, and vice versa. For example, statistic-based methods require thousands of distribution network for key statistical properties qualification, and simulated planning-based methods need a large amount of local geographical and social data. Therefore, we merely perform a qualitative comparison table shown in Table III, to illustrate the advantages of the proposed approach.

In addition, training time of UG-GAN and memory consumption are tested, to further evaluate the computational performance of the proposed method. The training time ranges from 1.4 to 1.8 seconds per iteration, with total time 4651 seconds for all 3000 iterations in this case. In terms of memory consumption, 1334 MB is used while training. In all, our method can be easily implemented on any standard computer with no additional configuration.

C. Application Examples

To further prove that the synthetic network generated by our model is realistic, a set of application examples are presented in this subsection.

1) Baseline Power Flow Analysis: Convergent AC power flow is the primary consideration to justify the network. Normally, when the load and system parameters are within reasonable limits, a converged AC power flow result can be obtained using the three-phase backward-forward algorithm [33]. The key point is to verify whether the voltage magnitude of each node is within the given limit (e.g., 0.95-1 p.u.). Fig. 12 illustrates the voltage magnitude of each node under the baseline power flow. In addition, the size of the solid circle represents the load of the node, and the thickness of the line represents the size of the line power. Noted that the voltage of bus #1 is assumed to be 1 p.u. in this case. It can be seen that the voltage of the generated system is within 0.966 p.u. to 1 p.u., which satisfies the voltage limits requirement. Besides, we also calculate the power flow on all overhead lines and underground cables based on the generated time series loads data, and they are all within the limit of chosen line conductors. Among them, the power flow on line 2-3 is the closest to the conductor limit in high load periods. Noted that, it still remains power flow margin.

2) Distributed Energy Resources Placement for Loss Reduction: In actual active distribution systems, DERs are also possibly installed by utilities for network loss reduction or renewable energy consumption. In this case, industrial load located at node #41 accounts for nearly two-thirds of the total load, and thus has a great influence on the total loss of this system. Thus, we try to vary the power flow of each line by installing DERs to reduce the loss. Based on the predetermined specific type of the DER [34], DERs placement is similar to the capacitor banks installation using the proposed MISCP formulation with minor modifications regarding the constraints on power injection. It is assumed that the total capacity of the DERs cannot be greater than the maximum load of the system. The results, including optimal sizes, locations, and the total amount of loss reduction, are shown in Table IV. Installing DER is not only an application of the generated active distribution network, but also expanding the scope of the application.

TABLE IV
RESULTS OF DISTRIBUTED ENERGY RESOURCES PLACEMENT

Case	Install			Total DERs	P_{loss} (kW)
No DER	-	-	-	0	1.911
Case1	Bus	40	-	1	0.303
	Size(kW)	1,216	-		
Case2	Bus	40	55	2	0.247
	Size(kW)	947	269		

3) Transmission and Distribution Power Flow Co-Analysis: As we discussed before, considering the unbalanced architecture of the distribution system, a zero-sequence current might be injected into the transmission system. Moreover, the characteristics of power flow are generally be changed with the installation of various components, such as capacitor banks and DERs, in current distribution networks. As a result, distribution network can no longer be directly regarded as an equivalent load of the transmission network. Transmission and distribution network time-series power flow co-analysis is important for ISO, using the detailed distribution network with similar properties. An application example is presented in this subsection.

The test system is obtained by replacing the aggregated load at bus 6 of the standard IEEE 9-bus transmission system with the generated distribution network, see Fig. 13. The test is carried out using Matlab and OpenDSS and the results are shown in Fig. 14. It can be observed that the voltage and power flow are within an acceptable range.

VI. CONCLUSION

This paper has proposed a deep learning-based framework to generate synthetic three-phase unbalanced active distribution networks using limited real data. Our method can implicitly capture the topological and electrical properties of real-world networks without revealing critical information. Moreover, the proposed method not only outputs grid connectivity but also effectively generates relevant time-series load data and locations and capacity of various grid components to obtain a comprehensive test case. With the proposed method, utilities will no longer have any concerns about making desensitized data publicly available at the request of industry and academia. Moreover, it is also possible for ISO of the transmission system to carry out transmission and distribution co-simulation based on generated networks for joint evaluation of the mutual effect of different systems. The results of case studies illustrate that these expectations can be met using the proposed method. Overall, our proposed method is able to consider the sparse network connectivity of the synthetic network merely by learning the distribution of biased random walks, and as a result, it greatly reduces the computation burden and improves the scalability at the stage of network synthesis. However, global convergence of the optimization problems may affect the scalability of our method, posed by large-scale distribution networks, in the network correction and extension process. Thus, the direction of our future research will focus on this issue, so as to extend our method to larger-scale distribution network synthesis task.

References

[1] M. H. Athari and Z. Wang, "Statistically characterizing the electrical parameters of the grid transformers and transmission lines," in *Proc. 10th Bulk Power Syst. Dyn. Control Symp.*, 2017, pp. 1–7.

[2] M. H. Athari and Z. Wang, "Introducing voltage-level dependent parameters to synthetic grid electrical topology," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4048–4056, Jul. 2019.

[3] A. Birchfield, K. Gegner, T. Xu, K. Shetye, and T. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1502–1510, Mar. 2017.

[4] S. Soltan and G. Zussman, "Generation of synthetic spatially embedded power grid networks," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2016, pp. 1–5.

[5] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3258–3265, Jul. 2017.

[6] D. Phillips, T. Xu, and T. Overbye, "Analysis of economic criteria in the creation of realistic synthetic power systems," in *Proc. IEEE Manchester PowerTech*, 2017, pp. 1–5.

[7] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 28–39, Jun. 2010.

[8] S. Soltan, A. Loh, and G. Zussman, "A learning-based method for generating synthetic power grids," *IEEE Syst. J.*, vol. 13, no. 1, pp. 625–634, Mar. 2019.

[9] M. Khodayar, J. Wang, and Z. Wang, "Deep generative graph distribution learning for synthetic power grids," 2019, *arXiv:1901.09674*. Q3

[10] A. Birchfield and T. Overbye, "Planning sensitivities for building contingency robustness and graph properties into large synthetic grids," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2020. Q4

[11] E. Schweitzer, A. Scaglione, A. Monti, and G. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 2, pp. 271–284, Jun. 2017.

[12] S. Abeysinghe, J. Wu, M. Sooriyabandara, M. Abeysekera, T. Xu, and C. Wang, "Topological properties of medium voltage electricity distribution networks," *Appl. Energy*, vol. 210, pp. 1101–1112, Jul. 2018. Q5

[13] S. Saha, E. Schweitzer, A. Scaglione, and N. Johnson, "A framework for generating synthetic distribution feeders using OpenStreetMap," in *Proc. 51st North Amer. Power Symp.*, 2019, pp. 1–6.

[14] A. Trpovski, D. Recalde, and T. Hamacher, "Synthetic distribution grid generation using power system planning: Case study of Singapore," in *Proc. 53rd Int. Univ. Power Eng. Conf.*, 2018, pp. 1–6.

[15] B. Palmintier et al., "Experiences developing large-scale synthetic U.S.-style distribution test systems," *Electric Power Syst. Res.* [Online]. Available: https://www.osti.gov/biblio/1770897

[16] K. Schneider, Y. Chen, D. Engle, and D. Chassin, "A taxonomy of North American radial distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2009, pp. 1–6.

[17] H. Li et al., "Building highly detailed synthetic electric grid data sets for combined transmission and distribution systems," *IEEE Open Access J. Power Energy*, vol. 7, pp. 478–488, 2020.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[19] J. Brownlee, *Generative Adversarial Networks With Python: Deep Learning Generative Models for Image Synthesis and Image Translation*, Mach. Learn. Mastery, 2019. Q6

[20] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3265–3275, May 2018.

[21] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "Outage detection in partially observable distribution systems using smart meters and generative adversarial networks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5418–5430, Nov. 2020.

[22] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann, "NetGAN: Generating graphs via random walks," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 609–618.

[23] Y. Guo, Y. Yuan, and Z. Wang, "Distribution grid modeling using smart meter data," *IEEE Trans. Power Syst.*, vol. 37, no. 3, pp. 1995–2004, May 2022.

[24] W. H. Kersting, *Distribution System Modeling and Analysis*. Boca Raton, FL, USA: CRC Press, 2017.

[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–12.

[26] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 855–864.

[27] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 1601–1608.

[28] J. Xiao, J. Lu, and X. Li, "Davies Bouldin index based hierarchical initialization K-means," *Intell. Data Anal.*, vol. 21, no. 6, pp. 1327–1338, Nov. 2017.

[29] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang, and Y. Guo, "Disaggregating customer-level behind-the-meter PV generation using smart meter data and solar exemplars," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5417–5427, Nov. 2021.

[30] M. Farivar and S. H. Low, "Branch flow model: Relaxations and convexification—Part I," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2554–2564, Aug. 2013.

[31] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, "A time-series distribution test system based on real utility data," in *Proc. 51st North Amer. Power Symp.*, 2019, pp. 1–6.

[32] H. Iba, "Random tree generation for genetic programming," in *Proc. Int. Conf. Parallel Problem Solving Nature*, 1996, pp. 144–153.

[33] N. Kishor, "Backward forward power flow for balanced / unbalanced networks," May 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/74985-backward-forward-power-flow-for-balanced-unbalanced-networks

[34] D. Q. Hung and N. Mithulananthan, "Multiple distributed generator placement in primary distribution networks for loss reduction," *IEEE Trans. Ind. Electron.*, vol. 60, no. 4, pp. 1700–1708, Apr. 2013.

Rong Yan (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from the School of Electrical Engineering, Wuhan University, Wuhan, China, in 2016, and the Ph.D. degree in electrical engineering from the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2021. Since July 2021, he has been an Engineer with Power Dispatching and Control Center, China Southern Power Grid Co., Ltd, Guangzhou, China. From 2019 to 2020, he was a Visiting Student with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. His research interest include the application of data-driven methods in power system stability analysis.

Yuxuan Yuan (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Iowa State University, Ames, IA, USA, in 2017, where he is currently working toward the Ph.D. degree. His research interests include distribution system state estimation, synthetic networks, data analytics, and machine learning.

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently Northrop Grumman Endowed Associate Professor with Iowa State University, Ames, IA, USA. He is also the Principal Investigator for a multitude of projects funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. His research interests include optimization and data analytics in power distribution systems and microgrids. He was the recipient of the National Science Foundation CAREER Award, Society-Level Outstanding Young Engineer Award from IEEE Power and Energy Society (PES), Northrop Grumman Endowment, College of Engineering's Early Achievement in Research Award, and the Harpole-Pentair Young Faculty Award Endowment. He is the Chair of IEEE PES PSOPE Award Subcommittee, Vice Chair of PES Distribution System Operation and Planning Subcommittee, and Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is the Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, IEEE POWER ENGINEERING LETTERS, and IET Smart Grid.

Guangchao Geng (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2009 and 2014, respectively. From 2012 to 2013, he was a Visiting Student with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From 2014 to 2017, he was a Postdoctoral fellow with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, and Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. He is currently an Associate Professor with the College of Electrical Engineering, Zhejiang University. His research interest includes power system stability and control, the applications of Internet of Things (IoT) technique and high performance computing (HPC) technique in power systems.

Quanyuan Jiang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1997, 2000, and 2003, respectively. From 2006 to 2008, he was a Visiting Associate Professor with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. He is currently a Professor with the College of Electrical Engineering and the Associate Dean of Undergraduate School, Zhejiang University, Hangzhou, China. His research interest includes power system stability and control, applications of energy storage systems, and high performance computing technique in power systems.