

Load Profile Inpainting for Missing Load Data Restoration and Baseline Estimation

Yiyan Li, *Member, IEEE*, Lidong Song, *Student Member, IEEE*, Yi Hu, *Student Member, IEEE*, Hanpyo Lee, *Student Member, IEEE*, Di Wu, *Member, IEEE*, PJ Rehm, and Ning Lu, *Fellow, IEEE*

Abstract—This paper introduces a Generative Adversarial Nets (GAN) based, *Load Profile Inpainting Network (Load-PIN)* for restoring missing load data segments and estimating the baseline for a demand response event. The inputs are time series load data before and after the inpainting period together with explanatory variables (e.g., weather data). We propose a Generator structure consisting of a coarse network and a fine-tuning network. The coarse network provides an initial estimation of the data segment in the inpainting period. The fine-tuning network consists of self-attention blocks and gated convolution layers for adjusting the initial estimations. Loss functions are specially designed for the fine-tuning and the discriminator networks to enhance both the point-to-point accuracy and realisticness of the results. We test the Load-PIN on three real-world data sets for two applications: patching missing data and deriving baselines of conservation voltage reduction (CVR) events. We benchmark the performance of Load-PIN with five existing deep-learning methods. Our simulation results show that, compared with the state-of-the-art methods, Load-PIN can handle varying-length missing data events and achieve 15-30% accuracy improvement.

Index Terms—Conservation voltage reduction, deep learning, Generative Adversarial Nets (GAN), gated convolution, generative adversarial network, missing data restoration, self-attention mechanism.

I. INTRODUCTION

MISSING data is a common issue in distribution system load profile processing. Often times, missing load data are caused by temporarily communication lost with the equipment. By counting the lengths of every missing load data segment in several real-world smart meter data sets provided by Pecan Street^[1] and utilities such as Duke Energy and Electricities, we found that approximately 70% of missing load data segments are less than 4 hours. Thus, restoring short missing load data segments is critical for improving data quality, which can further benefit downstream data-driven tasks, such as load forecasting, non-intrusive load monitoring, load pattern analysis, etc.

Moreover, the demand response (DR) baseline estimation problem can be considered as a special case of the missing load

data restoration problem, thus can be solved by directly implementing the missing load data restoration algorithm. This is because the DR baseline estimation is to recover the would-have-been load profiles assuming the DR program had not been executed, which is equivalent to restoring a missing load segment (i.e. the DR period) based on the normal load data before/after the DR period. For example, Conservation Voltage Reduction (CVR) is widely used by utilities for peak load reduction [1]. During a CVR event, system voltage at the substation bus will be reduced by 2-4% to achieve load reduction. To quantify the CVR-caused load reduction, it is very important for utility engineers to accurately estimate what the original load profile (i.e., the baseline) during a CVR event would have been had the bus voltage not been reduced. Because the pre- and post- CVR load profiles represent customer consumption patterns under normal system voltage, uncovering the CVR baseline is equivalent to restoring missing data in the CVR period. Because baseline estimation is essential to DR performance evaluation, inpainting the would-have-been load profile during a DR event is highly valuable to load service providers.

Existing missing data restoration methods for load profile inpainting are categorized into *model-based* and *data-driven*. Model-based methods use physical system models to simulate responses to external disturbances in hope of restoring missing data segments. For example, to estimate the CVR baseline, researchers use the distribution system topology and load models to predict load changes when system voltage changes [3]-[7]. In general, model-based methods require accurate distribution system models. However, in practice, distribution system models are incomplete and inaccurate due to partial observability and topology changes. Moreover, load composition varies with respect to the time-of-the-day while customer consumption patterns shift constantly due to occupancy and weather conditions. To tackle these challenges, [8] proposed a robust time-varying load modeling method to dynamically identify the parameters for the ZIP load model at the substation level, based on which the load-to-voltage dependence and the CVR factor can be estimated. [9]

This research is supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number DE-EE0008770.

Yiyan Li is with the College of Smart Energy, Shanghai Jiao Tong University, Shanghai Non-carbon energy conversion and utilization institute, Shanghai, 200240, China, and also with the Electrical and Computer Engineering Department, Future Renewable Energy Delivery and Management Systems Center, North Carolina State University, Raleigh, NC 27606 USA (email: yiyang.li@sjtu.edu.cn).

Lidong Song (corresponding), Yi Hu, Han pyo Lee, and Ning Lu are with the Electrical & Computer Engineering Department, Future Renewable Energy Delivery and Management (FREEDM) Systems Center, North Carolina State University, Raleigh, NC 27606 USA. (e-mails: lsong4@ncsu.edu, yhu28@ncsu.edu, hlee39@ncsu.edu, nlu2@ncsu.edu).

PJ Rehm is with Electricities (e-mail: prehm@electricities.org). Di Wu is with Pacific Northwest National Laboratory (e-mail: Di.Wu@pnnl.gov).

formulated the ZIP model parameter identification problem as an over-determined problem, and solved the problem dynamically to estimate the CVR factor using a sliding window method based on field measurements. [8][9] enhanced the applicability and accuracy of the model-based approaches.

Data-driven approach provides an end-to-end solution for missing data restoration, which is more convenient and has been widely studied. Data-driven methods can be further categorized into two approaches: *similarity-based* and *regression-based*. The similarity-based approach groups load profiles by day type, weather conditions, and shape characteristics of load profiles. The missing data segments are restored by referencing to the data on the load profiles having the best similarity match. Similarity-based methods are straightforward, easy to implement and explainable, therefore are widely used in field implementation [10]-[13]. However, in many cases, similarity metrics are normally defined by human analysts and can be based on the weighted average of many factors (e.g., weather, time, geographical conditions, and load types). This makes the accuracy of the method dependent on subjective selections of similarity metrics and weights.

Regression-based methods include linear regression [14], Long Short Term Memory (LSTM) [15], Stacked Autoencoder (SAE) [16], Gaussian Regression [17], Support Vector Regression (SVR) [18][19], etc. Regression-based methods usually achieve higher estimation accuracy compared to the similar day approach because of their nonlinear learning capabilities, especially when using deep-learning models. However, compared with similarity-based methods, the deep-learning based methods are less explainable and having higher computing costs. In recent years, hybrid solutions combining multiple regression models [20]-[22] are proposed for baseline estimation or missing data restoration.

The main drawbacks of traditional regression-based models are that the data format of the input and output is required to be fixed. However, in practice, the duration of missing data (model output) varies from minutes to several hours, and the length and number of available measurements (model input) also vary case by case. To cope with the varying-length cases, traditional methods need either increase the output window to cover the longest event or train separate models for different scenarios. This increases model complexity with additional computing and deployment costs, and the model performance can hardly be guaranteed due to information losses.

In this paper, we propose an advanced regression-based approach for missing data restoration by introducing the Generative Adversarial Nets (GAN). Studies of using GAN to solve the missing data restoration problem is still in infancy. In [23][24], the authors discuss the basic theory of GAN in restoring missing data, while in [25][26], the authors implement a GAN-based method in power domain to restore the grid measurement data and the PV profiles. Inspired by image inpainting, we develop a Load Profile Inpainting Network (Load-PIN) using GAN [27] as the basic structure of a highly accurate and flexible missing data restoration framework for recovering missing data on load profiles and estimating baselines for demand response events. The generator consists

of a coarse network and a fine-tuning network. Initially, the bidirectional time series load data before and after the missing data segment together with the explanatory variables are fed into the coarse network to obtain an initial estimation for the missing part. Next, initial estimations are sent to the fine-tuning network consisting of Gated Convolution layers [28] and Multi-head self-attention blocks [29] to improve accuracy. The generator network is trained under the guidance of the discriminator with specially designed loss functions.

The main contribution of this paper is the development of the Load-PIN model for missing data restoration and baseline estimation. The most distinct feature of the Load-PIN model is its flexibility in restoring variable-length data segments and its superior accuracy compared with the state-of-the-art methods. To the best of the authors' knowledge, the generative approach has not yet been developed for patching missing smart meter data and for demand response baseline identification.

The rest of the paper are organized as follows: Section II formulates the missing data restoration and CVR baseload estimation problem, and introduce the proposed Load-PIN model, Section III demonstrates the case study results, and Section IV concludes this paper.

II. METHODOLOGY

In this section, we first illustrate the problem formulation of load profile inpainting and the background of the GAN based approach. Then, we present the proposed Load-PIN framework including the 2-stage GAN generator structure and the loss function design.

A. Problem Formulation of Load Profile Inpainting

Denote a historical time series matrix of load as $\mathbf{Y} = [y_1, y_2, \dots, y_L]$, where L is the length of the time series. Denote the explanatory variables $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_L^1 \\ x_1^2 & x_2^2 & \dots & x_L^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^E & x_2^E & \dots & x_L^E \end{bmatrix} \quad (1)$$

where E is the number of explanatory variables.

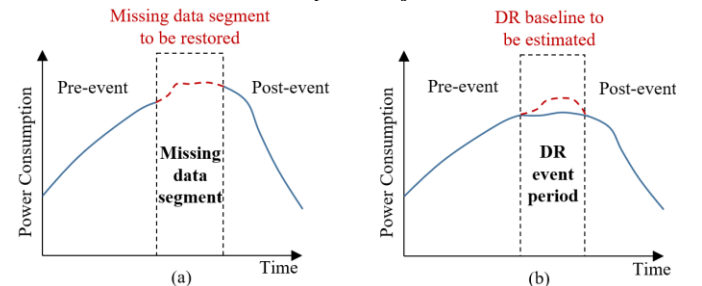


Fig. 1. Two basic event types: (a) missing data restoration, and (b) DR baseline identification. Blue solid lines are field measurements and red dotted lines are uncovered data segments.

Define an event to be a missing data segment (see Fig. 1(a)) or an unknown DR baseline (see Fig. 1(b)). Assume there are N events in $\mathbf{Y}$ and the duration of the i^{th} event (i.e. the length of the i^{th} inpainting period) is T_{event}^i

$$T_{event}^i = T_{end}^i - T_{start}^i \quad (2)$$

where T_{start}^i and T_{end}^i represents the starting and ending time stamps of the i^{th} event respectively, shown in Fig. 2. Meanwhile, the i^{th} restored data segment, $\hat{Y}_{event}^i$, is

$$\hat{Y}_{event}^i = [\hat{y}_{T_{start}^i}^i, \hat{y}_{T_{start}^i+1}^i, \dots, \hat{y}_{T_{end}^i}^i] \quad (3)$$

As shown in Fig. 2, we divide the load profile containing the i^{th} event into three periods: $[X_{event}^i, Y_{event}^i]$ as the inpainting data period, $[X_{pre}^i, Y_{pre}^i]$ as the pre-event period, and $[X_{post}^i, Y_{post}^i]$ as the post-event period.

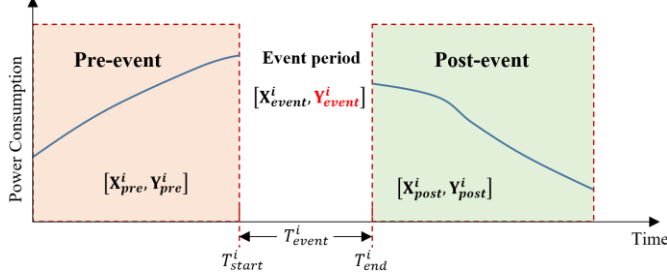


Fig. 2. An illustration of the division of the load profile containing an event.

Thus, a load profile inpainting problem can be described as

$$\hat{Y}_{event}^i = f_{\theta}(X_{pre}^i, Y_{pre}^i, X_{event}^i, X_{post}^i, Y_{post}^i) \quad (4)$$

where f_{θ} is the mapping function.

In the similarity-based approach, pre- and post- event data are mainly used for identifying similar days. Traditional forecasting-based methods use only pre-event data, X_{pre}^i and Y_{pre}^i , as model inputs to forecast $\hat{Y}_{event}^i$. Most correlation-based methods take X_{event}^i as inputs to build the mapping function between Y and X (e.g., temperature and voltage) for estimating $\hat{Y}_{event}^i$. Thus, one of the main drawbacks of the

state-of-the-art methods is that the information contained in all five available data sets (i.e., $[X_{pre}^i, Y_{pre}^i, X_{event}^i, X_{post}^i, Y_{post}^i]$) have not yet been fully utilized in one shot. To resolve this deficiency, we take the GAN based approach to use all 5 available data sets as inputs to f_{θ} for predicting $\hat{Y}_{event}^i$.

B. GAN-based Approach

As shown in Fig. 3, a GAN model consists of a generator network (G) and a discriminator network (D). The input of the generator is a latent vector z , often a Gaussian noise. The generated data, $G(z)$, along with the actual data, x , are then passed to discriminator D . The goal of D is to distinguish real data sets from the fake ones. The training of a GAN model is an iterative, adversarial process: G tries to generate samples $G(z)$ to fool D ; D learns to identify $G(z)$ from x by assigning greater probabilities to x and smaller ones to $G(z)$. As introduced in [27], this process is formulated as a minimax game

$$\min_G \max_D R(D, G) = \mathbb{E}_{x \sim p(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))] \quad (5)$$

where $R(D, G)$ is the reward function, $p(x)$ and $p(z)$ are the probability distributions of training data and latent vector, $\mathbb{E}$ is the expectation operator.

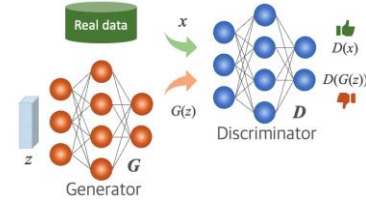


Fig. 3. The original GAN model.

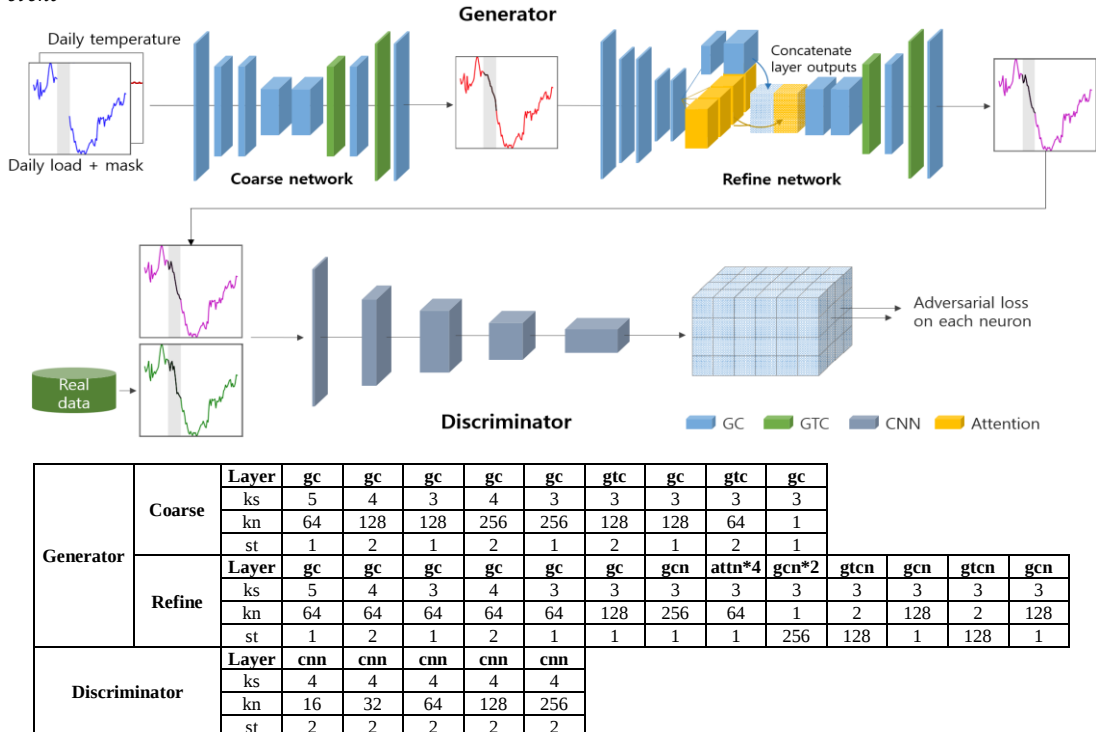


Fig. 4. The proposed Load-PIN framework. “GC” refers to a gated convolution block, “GTC” refers to a gated transpose convolution block, “CNN” refers to a convolutional block, and “Attention” refers to a self-attention block. “ks” means kernel size, “kn” means number of kernels, and “st” means stride.

C. Work Flow of the Load-PIN Framework

The Load-PIN framework is illustrated in Fig. 4. The model input $\mathbf{z}$ has three parts: 24-hour load and temperature profiles, and a Boolean mask indicating the event period as one and the normal period as zero. The load data resolution varies from 1-minute to 15-minute and the missing data duration, T_{event} , is less than 4 hours. The generator contains two stages: a coarse network for initial estimation and a fine-tuning network for polishing. The discriminator is a deep convolutional network with specially designed loss functions.

1) Training Sample Generation

The goal of the sample generation process is to generate samples evenly distributed in yearly load profiles so that the trained model will not be biased by factors such as the time-of-the-day and season-of-the-year. To achieve this goal, we slide a 24-hour moving window over the yearly historical load profiles, as shown in Fig. 5. In this paper, the time shift (Δt) of the moving window is one hour.

There are two considerations when creating training samples. *First*, to train and test the Load-PIN model, the training and testing samples should be generated from load profiles containing no CVR events. This is because, during a CVR event, system voltages are reduced by 2-4% for load reduction, making the load profile under the normal system voltage (the ground truth) unknown to the analyst to evaluate the model performance. *Second*, the missing data segment (indicated by the mask) will be positioned at the center of the 24-hour window so that the pre- and post- event data are equal in length.

After the Load-PIN model is trained, we apply it to CVR samples for CVR baseline identification. A CVR sample is a 24-hour load profile with the CVR event being placed in the middle of the 24-hour window (See Fig. 5).

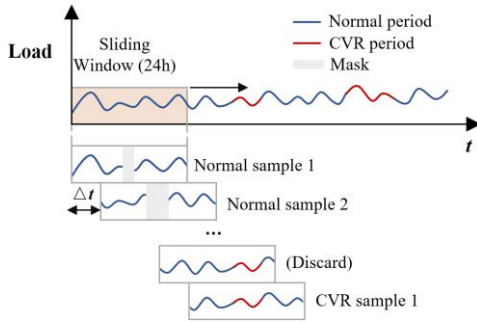


Fig. 5. Training sample generation process.

Denote N_{train} , N_{test} , and N_{CVR} as the total number of training samples, testing samples, and CVR samples, respectively. For the i^{th} normal sample, a mask of variable length (but less than 4-hour) is placed in the middle of the 24-hour window. The data being masked is the ground truth ($\mathbf{Y}_{event}^i$) of the forecasted missing data segment ($\hat{\mathbf{Y}}_{event}^i$). Before and after the mask are the pre-event and post-event load segments (i.e., $\mathbf{Y}_{pre}^i, \mathbf{Y}_{post}^i$). Meanwhile, load segments are paired with their corresponding temperature profiles as the explanatory variable (i.e., $\mathbf{X}_{pre}^i, \mathbf{X}_{event}^i, \mathbf{X}_{post}^i$). After $\hat{\mathbf{Y}}_{event}^i$ is estimated using (4), we can optimize the model parameters by

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^{N_{train}} \|\mathbf{Y}_{event}^i - \hat{\mathbf{Y}}_{event}^i\|_2^2 \quad (6)$$

2) Generator Network: 1st-Stage Coarse Network

In the first stage, we employ the Gated Convolution Network (GCN) [28] to formulate an encoder-decoder structure to restore the masked segments. Compared with conventional convolution neural network (CNN), GCN adds an additional feature-wise gating control mechanism, which learns soft masks automatically to compute the hidden layer features as

$$h_i(\mathbf{X}) = \varphi(\mathbf{X} \cdot \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X} \cdot \mathbf{U} + \mathbf{c}) \quad (7)$$

where $\mathbf{X}$ is the input of layer h_i , $\mathbf{W}$, $\mathbf{b}$, $\mathbf{U}$, $\mathbf{c}$ are trainable parameters. σ is the sigmoid function, φ can be any activation function (e.g., ReLU and LeakyReLU), and $\otimes$ is the element-wise product between matrices. This equation demonstrates that different weights of filters are applied at different temporal points to produce the output, resulting in a dynamic feature selection mechanism for each channel and each time point. Besides, Gated Transpose Convolution Network (GTCN) [30] layers are introduced to recover the estimated daily profiles from the GCN output.

The 1st-stage coarse network is trained purely based on the point-to-point content loss function, L_{coarse} .

$$L_{coarse} = \frac{1}{H} \|\mathbf{G}_{\theta_{G1}}(\mathbf{z}) - \mathbf{P}\|_2^2 \quad (8)$$

where θ_{G1} is the parameter of the coarse generator network, $\mathbf{P}$ is the ground truth load profile, and H is the dimension of $\mathbf{P}$.

Note that instead of computing the point-to-point content loss (L_{coarse}) for the entire 24-hour period, we calculate only the content loss for the masked segment plus a few points before and after the masked segment, the length of which is H . For example, in this paper, H is set as 5 hours to cover the masked segment with a margin of 0.5 hour. This ensures that the coarse network focuses on the masked period to enhance accuracy and achieve smoother transition between non-event periods and the event period (i.e., no spikes during the transitions).

3) Generator Network: 2nd-stage Fine-tuning Network

In the second stage, we make two modifications to the conventional GAN framework: adding *Multi-head self-attention* and *considering the tradeoff between content loss, feature matching loss and adversarial loss in the loss function* to polish the first-stage results and recover realistic details (high-frequency components).

Self-attention, also known as intra-attention, is a technique for focusing attention on various points in a single sequence when creating a representation of the sequence. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the value as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}\right)\mathbf{V} \quad (9)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ include learnable parameters representing query-key-value pairs, respectively, α is a scaling factor. Instead of performing a single attention function, [29] finds it beneficial to linearly project the queries, keys, and values

multiple times with different learned linear projections in parallel, which is multi-head attention formulated as

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_n) \mathbf{W}^O \quad (10)$$

$$\text{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V)$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ and $\mathbf{W}^O$ are the learnable parameter matrices of the projection. This allows the temporal dependencies between the elements in the input sequences to be modeled without considering their distances [31].

The loss function of the fine-tuning network (L_{refine}) includes 3 terms: the content loss, the adversarial loss (L_{adv}) and the feature-matching loss (L_{feat}), as shown in (11)-(13). λ_1 and λ_2 are the weights. Same as in L_{coarse} , the content loss minimizes point-to-point errors. L_{adv} improves the realisticness of the estimation results by maximizing the scores of the discriminator, where M is the dimension of the discriminator output (shown as the blue 3D matrix in Fig. 4). L_{feat} is defined in (13) as the distance between high-level feature maps extracted from the hidden layers of the discriminator network, where $\phi_j(\cdot)$ represents the output of the j^{th} intermediate convolution layer of the discriminator network. J is the number of intermediate layers in the discriminator network. Because high-level features of real load profiles are embedded in the hidden layer outputs, L_{feat} guides the fine-tuning network to generate more realistic results by matching those high-level features extracted from real profiles. Similar to L_{coarse} , all 3 loss terms in L_{refine} are calculated using the 5-hour segment instead of the 24-hour load profile.

$$L_{refine} = \frac{1}{H} \left\| G_{\theta_{G_2}}(\mathbf{z}) - \mathbf{P} \right\|_2^2 + \lambda_1 \cdot L_{adv} + \lambda_2 \cdot L_{feat} \quad (11)$$

$$L_{adv} = -\frac{1}{M} D(G_{\theta_{G_2}}(\mathbf{z})) \quad (12)$$

$$L_{feat} = \sum_{j=1}^J \left\| \phi_j(G_{\theta_{G_2}}(\mathbf{z})) - \phi_j(\mathbf{P}) \right\|_2^2 \quad (13)$$

4) Discriminator Network

The discriminator network is trained to solve the maximization problem defined by (5). In practice, an event can happen at any time of the day with varying lengths. To help the discriminator focus on the event duration, we also send the corresponding Boolean mask together with the load profile. The discriminator contains five convolutional layers with an increasing number of kernels. This allows us to compress the input profiles into high-level feature matrix, in which each element can cover the entire input load profile. Finally, the adversarial loss is applied to each neural to identify the fake and real inputs.

Inspired by [30], we adopt spectral normalization and hinge loss to stabilize the training process of the discriminator by minimizing the loss function L_D calculated as

$$L_D = \frac{1}{M} \text{ReLu}(1 - D_{\theta_D}(\mathbf{P})) + \frac{1}{M} \text{ReLU}(1 + D_{\theta_D}(G(\mathbf{z}))) \quad (14)$$

where θ_D is the parameters of the discriminator networks.

5) Model Performance Evaluation

Three performance metrics are calculated: normalized Root Mean Squared Error ($nRMSE$), Energy Error (EE), and $bias$.

$$nRMSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \sqrt{T_{event}^i \left\| \mathbf{Y}_{event}^i - \hat{\mathbf{Y}}_{event}^i \right\|_2^2} / \left\| \mathbf{Y}_{event}^i \right\|_1 \quad (15)$$

$$EE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left| \sum_{t=1}^{T_{event}^i} (y_t^i - \hat{y}_t^i) \right| / \left\| \mathbf{Y}_{event}^i \right\|_1 \quad (16)$$

$$bias = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(\frac{1}{T_{event}^i} \sum_{t=1}^{T_{event}^i} \frac{y_t^i - \hat{y}_t^i}{y_t^i} \right) \times 100\% \quad (17)$$

where $nRMSE$ evaluates the expected point-to-point error; EE evaluates the expected accumulated energy error; $bias$ reflects whether the model has a consistent difference between the actual and the missing data.

6) CVR Efficacy Estimation

To evaluate the performance of a CVR program containing N_{CVR} CVR events, the trained model is used to estimate the CVR baseline for each CVR sample so the expected load reduction when executing CVR can be calculated.

The raw average, normalized load reduction of the i^{th} CVR event is

$$CVR_{raw}^i = \frac{1}{T_{event}^i} \sum_{t=1}^{T_{event}^i} \frac{y_t^i - \hat{y}_t^i}{y_t^i} \times 100\% \quad (18)$$

The net average, normalized load reduction of the i^{th} CVR event considering the forecasting bias is

$$CVR_{net}^i = CVR_{raw}^i - bias \quad (19)$$

When calculating the forecasting bias in CVR baseline estimation, only the normal samples in the same season and having missing data during the CVR event periods are selected, i.e., the bias reflects the consistent difference between the actual and the missing data in CVR periods only.

If CVR_{net} is negative, load is reduced during the CVR period. If CVR_{net} is positive, load is increased during the CVR periods, making the feeder unfit for CVR.

The average CVR factor for all CVR events is

$$CVR_f = \frac{1}{N_{CVR}} \sum_{i=1}^{N_{CVR}} CVR_{net}^i / \Delta V^i \quad (20)$$

where ΔV^i is the voltage reduction ratio during the i^{th} CVR event. Note that the CVR factor can be used by utilities to identify suitable feeders for CVR.

III. CASE STUDY

In this paper, as shown in Table I, three test cases are set up to demonstrate the efficacy of the proposed method: Base case for performance benchmarking, fixed-duration CVR case, and variable-duration CVR case.

TABLE I
Test Case Descriptions

Case Description	Data Source	Data Resolution	Data Length	Data Size	CVR Events
1 Base	PECAN Street	1-minute	1-year	318 residential users	None

2	Fixed duration	Utility A	15-minute	2-year	3 feeders	24
3	Variable duration	Utility B	5-minute	1-year	1 substation	32

To benchmark Load-PIN, we compare its performance with five other deep learning models: Multi-layer Perceptron (MLP) [32], Long-short Term Memory (LSTM) [33], Temporal Convolutional Net (TCN) [34], Bi-LSTM (Bidirectional LSTM) [35], and SAE (Stacked Auto-Encoder) [36].

Using only the pre-event data (i.e., $\mathbf{X}_{pre}$ and $\mathbf{Y}_{pre}$) as inputs, MLP, TCN, and LSTM formulate missing data restoration and CVR baseline identification as a load forecasting problem. Bi-LSTM and SAE are bi-direction models. Thus, the inputs for the Bi-LSTM, SAE, and Load-PIN models are the same. Model hyper-parameters are selected based on the trial-and-error method.

In each case, all models are trained (70%), validated (15%) and tested (15%) on non-CVR samples. For the two CVR cases, the trained model is applied to the CVR samples for CVR baseline identification. The baselines are then used for computing CVR effects, which will be used to assess the efficacy of the CVR program.

A. Performance Evaluation for Missing Data Restoration

As shown in Table I, in the base case, PECAN Street data [1] is used for performance benchmarking on missing data restoration. To test the impact of data resolution on estimation accuracy, 1-minute data are down-sampled to 5-min, 15-min, 30-min and 1-h. To benchmark the model performance at different load aggregation levels, we test on aggregated load profiles of 10 users, 50 users, 100 users, 200 users and 300 users. After combining 5 data resolutions and 5 load levels, we obtain 25 test scenarios in total. The mask length is set to be 3-hour.

From the results, the following observations are made:

- As shown in Table II, the three bidirectional models (i.e., Bi-LSTM, SAE, and Load-PIN) outperform the one-directional forecasting models in most cases. As shown in Fig. 6, the bidirectional models (solid lines) has lower RMSEs than the forecasting models (dashed lines), especially in the second half of the 3-hour estimation window. The results clearly demonstrate that, by adding post-event data into the input, the accuracy will improve significantly. The step-by-step RMSE distributions shown in Fig. 7 further confirm the above observation.
- As shown in Table II, when data granularity is 5-min and 15-min, Load-PIN outperforms all other models and shows 15-30% improvement compared with the second-best model. This shows that Load-PIN can extract information hidden inside the high-resolution data for forecasting the missing data segments. However, if the data resolution is too low, the Load-PIN does not show significant performance improvements. This is because in those cases, forecasting average values outweigh uncovering load shape details.
- In order to demonstrate the effectiveness of the

proposed two-stage generator network in Load-PIN, we compare the Load-PIN performance with a vanilla-GAN model (i.e. remove the refine network from Load-PIN) to formulate a control experiment. Results are shown in Table II. We can see that the vanilla-GAN model has equal or even slightly better performance than Load-PIN on low-resolution test cases (30-min and 1-h), but much worse on high-resolution test cases (15-min, 5-min and 1-min). This result demonstrates the advantage of the proposed two-stage generator structure in capturing and recovering high-frequency information in the time series data. Because the most commonly-used sampling intervals for load data are 5-min or 15-min, the proposed Load-PIN model is recommended

- As shown in Fig. 8, when the load aggregation level increases from 10 to 300 users, both $nRMSE$ and EE decrease. As shown in Fig. 8(a), to achieve a smaller $nRMSE$, the best data granularity (highlighted by yellow rectangles) for a group with less than 100 users is 1-h and for 300 users is 15-min. This is because when load aggregation level increases, the load profile becomes smoother. Thus, it is possible to use higher data granularity for restoring more detailed load shapes, which, in turn, improves the point-to-point accuracy. According to Fig. 8(b), to achieve a smaller EE , the best data granularity (highlighted by yellow rectangles) is 15-minute.
- As shown in Fig. 9, Load-PIN performance is consistent across different hours of the day and in different seasons. Note that the error is substantially smaller for summer peak load periods (i.e. between 15:00-18:00 in summer months), which is highly desirable for CVR baseline identification.
- As shown in Fig. 10, we tested the Load-PIN model under different training sample sizes on 15-min, 300 user Pecan Street test case. Results show that the model performance is insensitive to the training sample size when the sample size is larger than 24 hours, which demonstrate the robustness of the Load-PIN model.

To summarize, Load-PIN outperforms all other algorithms in missing data restoration. Higher resolution data is preferred when there are more users in the group. In most cases, 5 or 15 minutes data set are sufficient for restoring a missing data segment with a duration of 4-hour or less.

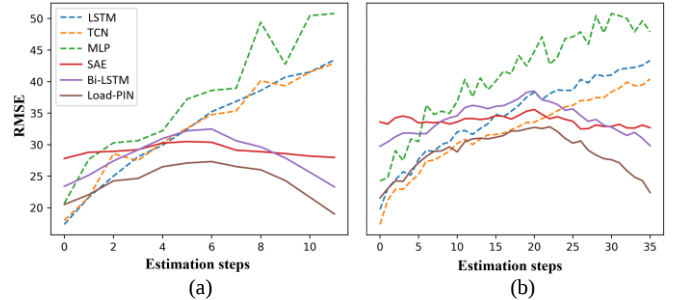


Fig. 6. Load-PIN performance evaluation: averaged step-by-step RMSE of the 6 models during 3-hour estimation interval. Test on the aggregated load of 300 users. (a) 15-min data granularity, (b) 5-min data granularity.

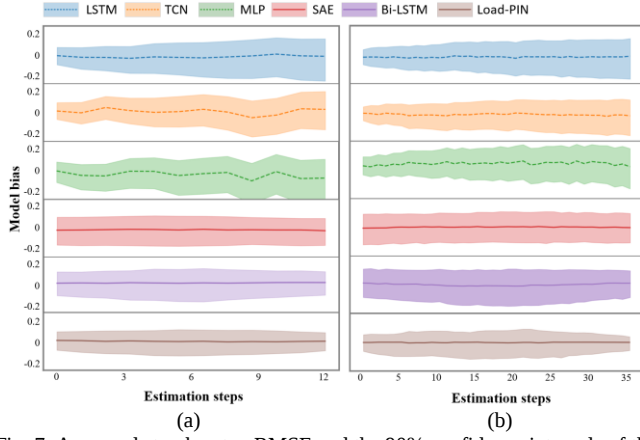


Fig. 7. Averaged step-by-step RMSE and the 90% confidence intervals of the 6 models during 3-hour estimation period. Test on the aggregated load of 300 users. (a) 15-min data granularity, (b) 5-min data granularity.

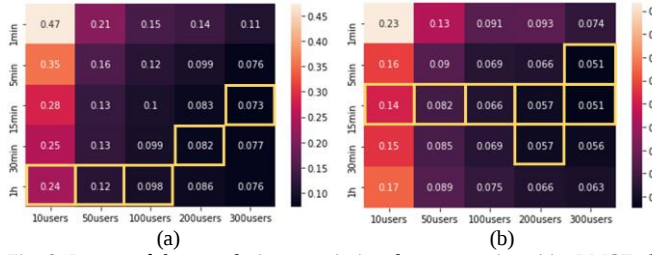


Fig. 8. Impact of data resolution on missing data restoration. (a) nRMSE, (b) EE. Note the value is an averaged of all 6 deep learning models. The yellow rectangles highlight the smallest error under each aggregation level.

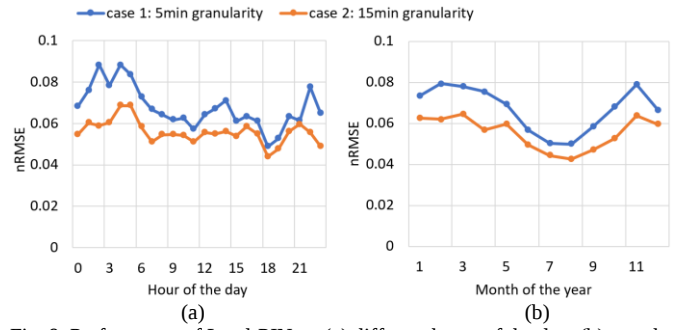


Fig. 9. Performance of Load-PIN on (a) different hours of the day, (b) months of the year.

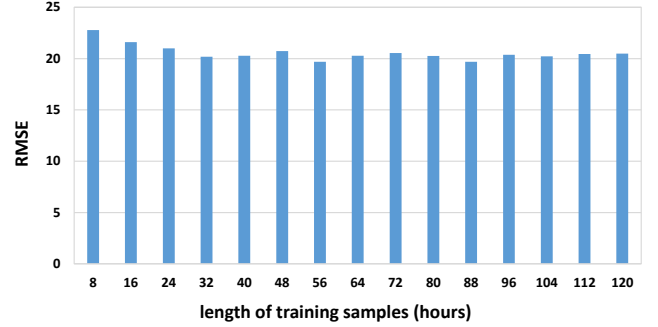


Fig. 10. Load-PIN performance under different size of training samples, 15-min, 300 user test case.

TABLE II
Model performances on the Pecan Street Test Case (Mask length is set to be 3 hours)

data granularity	aggregation level	nRMSE								EE							
		LSTM	TCN	MLP	SAE	Bi-LSTM	Vanilla-GAN	Load-PIN	Improve-ment	LSTM	TCN	MLP	SAE	Bi-LSTM	Vanilla-GAN	Load-PIN	Improve-ment
1-min	10users	0.46	0.43	0.39	0.35	0.57	0.54	0.35	-0.85%	0.22	0.21	0.16	0.15	0.26	0.32	0.15	-1.99%
	50users	0.25	0.21	0.19	0.15	0.22	0.26	0.12	18.90%	0.17	0.13	0.11	0.08	0.13	0.19	0.06	23.35%
	100users	0.19	0.17	0.15	0.11	0.16	0.25	0.10	7.79%	0.13	0.10	0.09	0.06	0.10	0.19	0.06	4.48%
	200users	0.16	0.15	0.12	0.08	0.12	0.23	0.07	15.74%	0.12	0.10	0.06	0.05	0.09	0.17	0.05	4.14%
	300users	0.15	0.13	0.10	0.07	0.11	0.22	0.09	-25.30%	0.12	0.09	0.06	0.05	0.08	0.17	0.05	-12.78%
	mean	0.24	0.22	0.19	0.15	0.24	0.30	0.15	3.26%	0.15	0.13	0.10	0.08	0.13	0.21	0.08	3.44%
5-min	10users	0.39	0.34	0.35	0.32	0.38	0.40	0.19	39.76%	0.20	0.13	0.15	0.15	0.17	0.15	0.08	38.46%
	50users	0.20	0.15	0.15	0.14	0.15	0.20	0.10	26.63%	0.14	0.08	0.08	0.08	0.08	0.12	0.07	9.76%
	100users	0.15	0.11	0.12	0.10	0.11	0.16	0.11	-7.36%	0.10	0.06	0.07	0.06	0.06	0.11	0.05	11.79%
	200users	0.15	0.09	0.12	0.08	0.08	0.13	0.07	17.36%	0.11	0.05	0.09	0.05	0.05	0.10	0.05	3.40%
	300users	0.08	0.08	0.10	0.07	0.07	0.12	0.06	13.73%	0.05	0.05	0.07	0.05	0.05	0.10	0.04	24.77%
	mean	0.19	0.15	0.17	0.14	0.16	0.20	0.11	18.02%	0.12	0.07	0.09	0.08	0.08	0.12	0.06	17.64%
15-min	10users	0.30	0.30	0.30	0.27	0.28	0.27	0.21	21.20%	0.15	0.15	0.15	0.14	0.15	0.12	0.09	26.74%
	50users	0.14	0.14	0.15	0.12	0.12	0.12	0.09	24.88%	0.09	0.10	0.09	0.08	0.07	0.07	0.05	25.60%
	100users	0.11	0.11	0.12	0.09	0.09	0.09	0.08	8.68%	0.07	0.07	0.09	0.06	0.06	0.05	0.05	7.59%
	200users	0.10	0.10	0.09	0.07	0.07	0.07	0.05	31.02%	0.07	0.07	0.06	0.05	0.05	0.04	0.04	5.32%
	300users	0.08	0.08	0.09	0.06	0.06	0.06	0.05	15.70%	0.06	0.05	0.07	0.05	0.05	0.04	0.03	21.10%
	mean	0.15	0.15	0.15	0.12	0.13	0.12	0.10	20.30%	0.09	0.09	0.09	0.08	0.07	0.06	0.05	22.94%
30-min	10users	0.27	0.27	0.28	0.23	0.23	0.21	0.22	-5.73%	0.15	0.15	0.19	0.15	0.14	0.12	0.11	5.41%
	50users	0.15	0.14	0.15	0.11	0.11	0.10	0.09	10.41%	0.12	0.09	0.09	0.08	0.07	0.06	0.07	-16.13%
	100users	0.11	0.12	0.12	0.08	0.08	0.07	0.08	-7.18%	0.08	0.09	0.08	0.06	0.06	0.05	0.05	-9.32%
	200users	0.10	0.10	0.10	0.07	0.07	0.06	0.06	-5.11%	0.07	0.06	0.07	0.05	0.05	0.04	0.04	-4.05%
	300users	0.09	0.10	0.10	0.06	0.06	0.05	0.05	-2.46%	0.06	0.06	0.07	0.05	0.05	0.03	0.04	-25.85%
	mean	0.14	0.15	0.15	0.11	0.11	0.10	0.10	-2.24%	0.10	0.09	0.10	0.08	0.07	0.06	0.06	-2.55%
1-h	10users	0.29	0.28	0.27	0.19	0.19	0.18	0.19	-3.45%	0.19	0.23	0.18	0.15	0.15	0.11	0.12	-5.79%
	50users	0.14	0.14	0.15	0.09	0.09	0.09	0.10	-12.00%	0.11	0.10	0.11	0.07	0.07	0.06	0.08	-31.52%
	100users	0.13	0.13	0.11	0.07	0.07	0.06	0.07	-8.64%	0.10	0.09	0.08	0.06	0.06	0.04	0.05	-11.12%
	200users	0.12	0.11	0.11	0.06	0.06	0.05	0.06	-16.91%	0.10	0.07	0.08	0.05	0.05	0.04	0.05	-36.27%
	300users	0.11	0.09	0.09	0.05	0.06	0.04	0.04	4.50%	0.09	0.06	0.07	0.04	0.05	0.03	0.04	-26.57%
	mean	0.16	0.15	0.15	0.09	0.09	0.09	0.09	-4.51%	0.12	0.11	0.10	0.07	0.08	0.06	0.07	-21.72%

B. Performance Evaluation on CVR Baseline Estimation

In this section, we apply the proposed Load-PIN model to estimate the CVR baseline for actual feeders in North Carolina, USA. For each feeder, we first train and test the Load-PIN model using data collected in non-CVR days (same as in Section III.A). Next, the trained model is used for baseline estimation in CVR days. The estimated baselines are used for CVR efficacy assessment.

1) Fixed-duration CVR Case

As shown in Table 1, the fixed CVR duration case is conducted using data collected from three residential distribution feeders in utility A, namely BR, DF and SL. All 24 labeled CVR events are in summer months with a fixed duration of 3 hours between 14:00 and 19:00 p.m. The CVR voltage reduction is 4%. The model performance is evaluated on non-CVR days. Note that forecasting bias is removed using methods introduced in Section II.C.6 when calculating CVR load reduction.

TABLE III
MODEL PERFORMANCES ON THE 3 TEST FEEDERS (IN PERCENTAGE)

(%)	Feeder	TCN	LSTM	MLP	SAE	BLSTM	Load-PIN
nRMSE	BR	3.12	3.13	3.8	4.05	3.67	2.50
	DF	3.87	3.78	5.56	6.35	5.8	2.86
	SL	3.66	3.72	3.91	4.03	3.67	3.02
EE	BR	2.02	2.03	2.63	2.91	2.54	1.13
	DF	2.86	2.75	4.03	4.91	3.78	1.98
	SL	2.54	2.65	2.77	2.47	2.18	1.40
Absolute Bias	BR	0.33	0.37	1.52	1.79	0.39	0.03
	DF	1.63	0.18	2.34	3.32	1.13	0.22
	SL	1.03	1.21	1.1	1.31	0.11	0.05

As shown in Table III, Load-PIN outperforms the other methods (lowest *nRMSE* and *EE*) by a large margin. From the results, we made the following observations

- As shown in Fig. 11, the results obtained by Load-PIN (the solid brown line) show similar trends as the average of all six models (the solid pink line). This shows that Load-PIN captures the trending information well.
- In Fig. 12, we randomly plot four CVR baselines for each feeder (out of 24 CVR events). The forecasted baseline shows a smooth transition from the pre-CVR to CVR periods and from the CVR to post-CVR periods.
- As shown in Figs. 11 and 12, for all three feeders, we observe clear load reduction (i.e. the value of CVR effect is negative in Fig. 11) in the first 1.5-hour. However, for feeders BR and DF, a pay-back period is observed in the second 1.5-hour. In fact, the amount of load reduction starts to diminish after the first hour. In some cases, feeder loads start to rise after the initial drop until exceeding the baseline.
- This CVR diminishing and subsequent pay-back effect is very likely caused by the increasing penetration of thermostatically controlled appliances (e.g., refrigerators, ovens, air conditioners) and LED lighting loads, which are rapidly replacing the incandescent lighting loads. Note that when voltage is low, appliances will turn on longer if a fixed amount of energy is required in each duty cycle.

- For feeder SL, even though the pay-back effect is less visible compared to feeders BR and DF, the CVR effect also diminishes after one hour.
- Overall, the results show that CVR efficacy will diminish in a prolonged CVR event, which indicates that the utilities may need to execute CVR for a period less than 2-hour.

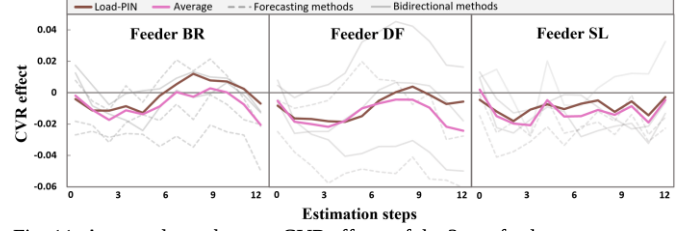


Fig. 11. Averaged step-by-step CVR effects of the 3 test feeders.

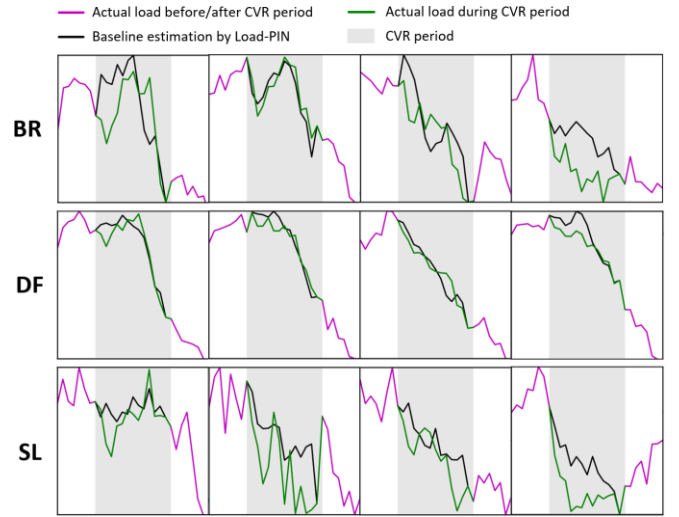


Fig. 12. Examples of the Load-PIN generated CVR baseline.

2) Variable-duration CVR Case

As shown in Table I, in this case, we use 5-minute data collected by a utility at a distribution substation bus to identify the CVR baseline for CVR events with variable durations. There are 32 CVR events in summer with durations ranging from 75 minutes up to 190 minutes. Figure 13 shows the daily voltage profiles at the feeder head of 3 identified CVR event days as an example.

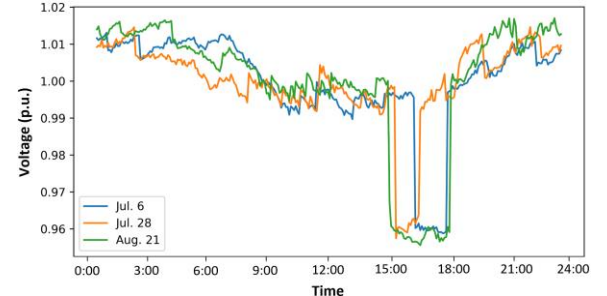


Fig. 13. Examples of daily voltage profiles of 3 CVR event days measured at the feeder head.

To make the bidirectional models (including Load-PIN) adapt to the varying-length CVR events, we randomly put a mask to each training sample with varying lengths between 1 - 4 hours. For the 3 forecasting models, we fix the output window to 4 hours so all possible CVR durations are covered in their

forecasting range.

As shown in Table IV, Load-PIN outperforms all other models by achieving the smallest $nRMSE$ and EE . The model bias is 0.43 only slightly higher than that of LSTM. The results demonstrate that Load-PIN can handle varying-length estimation tasks by leveraging the closest bidirectional data around each CVR events.

TABLE IV

MODEL PERFORMANCE ON VARYING-LENGTH FEEDER CASE (IN PERCENTAGE)

(%)	TCN	LSTM	MLP	SAE	BLSTM	Load-PIN
nRMSE	3.23	4.13	4.93	4.31	3.18	2.15
Energy	1.91	2.69	3.37	3.03	1.95	1.07
Absolute Bias	0.82	0.27	1.17	1.74	1.57	0.43

We implement the 6 trained models to estimate the CVR baseline for all the 32 events, and then calculate the averaged CVR effect, as shown in Fig. 14. Note that due to the uneven durations of the CVR events, results are less stable after 2h because the durations of many CVR events are less than 2 hours. However, we can still see similar trends with the fixed-duration case (using data from utility B): the CVR effect is more significant in the first hour and the initial load drop period will be followed by a pay-back period where load will rise.

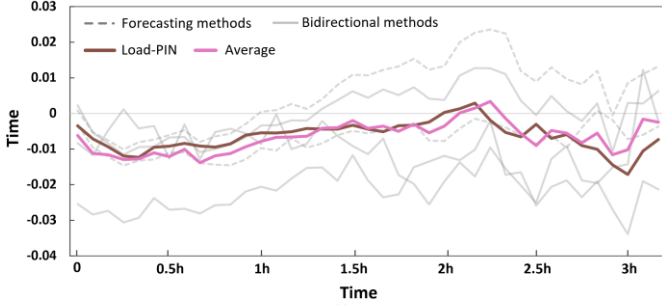


Fig. 14. Averaged step-by-step CVR effects at a substation bus.

IV. CONCLUSION

In this paper, we propose a novel deep-learning model, Load-PIN, to solve the missing data restoration and CVR baseload estimation problem. Load-PIN is a GAN-based framework including a two-stage generator and specially-designed loss functions that can solve the missing data restoration problem as a 2D image inpainting process. Gated Convolution and Multi-head self-attention mechanisms are leveraged in the framework to enhance the estimation accuracy. Load-PIN is trained using the dynamic-masking strategy so that it can handle CVR events with varying durations. Based on the Pecan Street test case, we demonstrate that at higher load aggregation levels, higher data resolution can achieve better estimation accuracy. In general, 5-min and 15-min resolutions are sufficient for feeder level studies. Meanwhile, we observe 15-30% accuracy improvement of the Load-PIN model under the suggested data granularity, compared with 5 benchmarking methods. This is because Load-PIN can extract features from both the bidirectional time series data and the explanatory variables to recover missing segments with realistic high-frequency details. Further experiments on different time of the day and different months of the year show that the Load-PIN model has stable performance along the day, and has superior

performance during the peak load periods in summer.

Using the trained Load-PIN model for CVR baseline identification, we computed the averaged CVR effects for CVR programs with fixed- and variable- durations based on the labeled feeder-level data provided by local utilities. We show that CVR can achieve load reduction in the first 1 hour. However, after 1.5 hours, the CVR effect starts to diminish and a pay-back period can be observed. This may cause unexpected load peaks in post-CVR periods. From the results, we want to make two recommendations. First, the CVR execution duration should be less than 2-hours. Second, feeders with high penetration of thermostatically controlled loads may not be good candidates for prolonged CVR programs.

Future work may focus on introducing transfer learning to the framework to evaluate its generality across different data sets. Also, we will try to build an online version of the Load-PIN model to give a real-time effect estimation for ongoing CVR events.

REFERENCES

- [1] Pecan Street. Available at: <https://www.pecanstreet.org/>
- [2] Wang, Zhaoyu, and Jianhui Wang. "Review on implementation and assessment of conservation voltage reduction." *IEEE Transactions on Power Systems* 29.3 (2013): 1306-1315.
- [3] Visconti, Igor F., Delberis A. Lima, and Jovica V. Milanović. "Comprehensive analysis of conservation voltage reduction: A real case study." 2019 IEEE Milan PowerTech. IEEE, 2019.
- [4] Zhang, Yongxi, et al. "Optimal placement of battery energy storage in distribution networks considering conservation voltage reduction and stochastic load composition." *IET Generation, Transmission & Distribution* 11.15 (2017): 3862-3870.
- [5] Wang, Zhaoyu, and Jianhui Wang. "Time-varying stochastic assessment of conservation voltage reduction based on load modeling." *IEEE Transactions on Power Systems* 29.5 (2014): 2321-2328.
- [6] Diaz-Aguiló, Marc, et al. "Field-validated load model for the analysis of CVR in distribution secondary networks: Energy conservation." *IEEE Transactions on Power Delivery* 28.4 (2013): 2428-2436.
- [7] Schneider, Kevin P., et al. Evaluation of conservation voltage reduction (CVR) on a national level. No. PNNL-19596. Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2010.
- [8] Zhao J, Wang Z, Wang J. Robust time-varying load modeling for conservation voltage reduction assessment[J]. *IEEE Transactions on Smart Grid*, 2016, 9(4): 3304-3312.
- [9] Ma Z, Xiang Y, Wang Z. Robust Conservation Voltage Reduction Evaluation Using Soft Constrained Gradient Analysis[J]. *IEEE Transactions on Power Systems*, 2022, 37(6): 4485-4496.
- [10] Coughlin, Katie, et al. Estimating demand and response load impacts: Evaluation of baselineload models for non-residential buildings in california. No. LBNL-63728. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2008.
- [11] Xiang, Biao, et al. "Smart Households' Available Aggregated Capacity Day-ahead Forecast Model for Load Aggregators under Incentive-based Demand Response Program." 2019 IEEE Industry Applications Society Annual Meeting. IEEE, 2019.
- [12] Wijaya, Tri Kurniawan, Matteo Vasirani, and Karl Aberer. "When bias matters: An economic assessment of demand response baselines for residential customers." *IEEE Transactions on Smart Grid* 5.4 (2014): 1755-1763.
- [13] Han Pyo Lee, Lidong Song, Yiyang Li, Ning Lu, Di Wu, PJ Rehm, Matthew Makdad, Edmond Miller, "An Iterative Bidirectional Gradient Boosting Algorithm for CVR Baseline Estimation" 23PESGM0022, submitted to 2023 IEEE PES General Meeting, Available online at: <http://arxiv.org/abs/2211.03733>
- [14] Matsukawa, Shun, et al. "Stable segment method for multiple linear regression on baseline estimation for smart grid fast automated demand response." 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia). IEEE, 2019.
- [15] Oyedokun, James, et al. "Customer baseline load estimation for incentive-based demand response using long short-term memory recurrent neural

- network." 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe). IEEE, 2019.
- [16] Chen, Yang, et al. "Privacy-Preserving Baseline Load Reconstruction for Residential Demand Response Considering Distributed Energy Resources." *IEEE Transactions on Industrial Informatics* (2021).
 - [17] Weng, Yang, Jiafan Yu, and Ram Rajagopal. "Probabilistic baseline estimation based on load patterns for better residential customer rewards." *International Journal of Electrical Power & Energy Systems* 100 (2018): 508-516.
 - [18] Chen, Yongbao, et al. "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings." *Applied Energy* 195 (2017): 659-670.
 - [19] Wang, Zhaoyu, Miroslav Begovic, and Jianhui Wang. "Analysis of conservation voltage reduction effects based on multistage SVR and stochastic process." *IEEE Transactions on Smart Grid* 5.1 (2013): 431-439.
 - [20] Sun, Mingyang, et al. "Clustering-based residential baseline estimation: A probabilistic perspective." *IEEE Transactions on Smart Grid* 10.6 (2019): 6014-6028.
 - [21] Zhang, Yufan, Qian Ai, and Zhaoyu Li. "Improving aggregated baseline load estimation by Gaussian mixture model." *Energy Reports* 6 (2020): 1221-1225.
 - [22] Ge, Xinxin, et al. "Spatio-Temporal Two-Dimensions Data Based Customer Baseline Load Estimation Approach Using LASSO Regression." *IEEE Transactions on Industry Applications* (2022).
 - [23] Yoon, Jinsung, James Jordon, and Mihaela Schaar. "Gain: Missing data imputation using generative adversarial nets." *International conference on machine learning*. PMLR, 2018.
 - [24] Luo, Yonghong, et al. "Multivariate time series imputation with generative adversarial networks." *Advances in neural information processing systems* 31 (2018).
 - [25] Zhang, Kexin, Xiaobo Dou, and Xiaolong Xiao. "Grid Defect Data Completion Based on Generative Adversarial Imputation Nets." 2021 IEEE Sustainable Power and Energy Conference (iSPEC). IEEE, 2021.
 - [26] Zhang, Wenjie, et al. "SolarGAN: Multivariate solar data imputation using generative adversarial network." *IEEE Transactions on Sustainable Energy* 12.1 (2020): 743-746.
 - [27] I. Goodfellow et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672-2680, 2014.
 - [28] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, 2017: PMLR, pp. 933-941.
 - [29] A. Vaswani et al., "Attention is all you need," vol. 30, 2017.
 - [30] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471-4480.
 - [31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
 - [32] Dudek, Grzegorz. "Multilayer perceptron for short-term load forecasting: from global to local approach." *Neural Computing and Applications* 32.8 (2020): 3695-3707.
 - [33] Zheng, Jian, et al. "Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network." 2017 51st Annual conference on information sciences and systems (CISS). IEEE, 2017.
 - [34] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." *arXiv preprint arXiv:1803.01271* (2018).
 - [35] Graves, Alex, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM." 2013 IEEE workshop on automatic speech recognition and understanding. IEEE, 2013.
 - [36] Gehring, Jonas, et al. "Extracting deep bottleneck features using stacked auto-encoders." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.