

Enhance Unobservable Solar Generation Estimation via Constructive Generative Adversarial Networks

Jingyi Yuan, *Student Member, IEEE* and Yang Weng, *Senior Member, IEEE*

Abstract—Power distribution grids experience the proliferation of solar photovoltaics (PV) at the system edge. However, its counterpart of sparse meter deployment provides insufficient monitoring of PVs, for which the potential violations challenge the operators for energy management and stable operation. Some previous works use satellite imagery to detect distributed PVs for easy access to data. However, their PV localization methods rely on the label-rich area with a unitary background/environment to implement well; even further/harder, they do not provide precise metered-PV detection and quantification to estimate/know PV generation outputs in unobservable areas, which is essential to prevent the edge from excessive two-way power flow and other violations. Thus, we combine the two steps of detecting PV existence and quantifying PV amount into one classification task. To boost the classification performance in the unobservable edge area, we construct a generative adversarial network that simultaneously augments the diversity of labeled PV satellite images and embeds distinct PV characteristics/features for training the classifier. Furthermore, the PV localization and quantification result is combined with geographic information, historical weather conditions, and neighboring generation patterns to estimate power output at the system edge. We validate the proposed approaches on PV systems in the southwest of the U.S. Experiment results show high accuracy and robustness in predicting distributed solar power without sufficient prior information.

Index Terms—Distributed PV forecast, diversified data sources, structured learning, weak supervised learning, variational GANs.

I. INTRODUCTION

Sustainable and inexhaustible solar energy is one of the fastest increasing renewable resources in the smart power grid. For example, research estimates that 150 - 530 GW of cumulative solar-based power will potentially be available in the U.S. by 2040 [1]. Unlike the conventional energy source with scheduled power output, PV generation depends on various temporal and spatial factors, e.g., weather, atmospheric conditions, and installation position and quantity [2]. They naturally bring variability and uncertainty, leading to bidirectional power flow and frequent voltage fluctuations of voltages and currents in distribution grids [3], [4]. To maintain the safety and reliability of the grid, distribution system operators (DSOs) require accurate information on the solar panel locations and PV generation forecasts for system remodeling and predictive energy management [5]–[7]. Moreover, the foreseeable future of PV energy-sharing and its economics posits in urban areas raise a high demand for precise and easily accessible information on distributed PV systems [8], [9].

For PV generation forecast, existing approaches can be divided into two folds: the physical model-based methods and the data-driven approaches [10]–[20]. Physical model-based methods rely on numerical weather prediction (NWP)

or satellite imagery to analyze the atmospheric conditions for solar irradiance, with which the power output is computed using physical characteristics [10], [13]. Some of the data-driven approaches directly estimate PV outputs from historical data, primarily extracting the statistical properties from PV measurements for new predictions [11], [12]. Machine learning models have been developed to consider highly correlated factors to characterize solar irradiance and predict PV generation. Such methods require complete information on PV locations, panel numbers, solar meter measurements, etc. However, in power distribution grids, the scattered PV generation data may come from different sources, including solar panel/inverter manufacturers, PV system development companies, utilities, and residential/commercial consumers. The methods mentioned above rely on the timely data aggregation from these different sources, which requires constant and intensive manual efforts [21], [22]. Thus, the knowledge is often incomplete or unavailable, especially for residential-level consumers. For example, the National Renewable Energy Laboratory (NREL) focuses on an Open PV Project to track distributed PV installations. The project relies on voluntary surveys and self-reports to provide a general understanding of PV distribution but is still unreliable in precision. Moreover, the database is easily outdated due to the rapid growth and widespread PV installations [23].

Instead of multiple PV data sources, the other group of data-driven methods uses smart meter data to detect unauthorized PV installation, and estimate behind-the-meter PV generation [14]–[20]. These methods are summarized in Fig. 1. Specifically, smart meters record the net load data of customers, which is the conventional load minus the PV generation. While the PV is invisible, different model-free methods are adopted to uncover its existence and generation output. For example, the change points in historical measurements are detected, verified, and estimated, but the unsupervised method requires predefined hyperparameters to work properly [14]. The supervised learning methods in [17], [20] need sufficient labeled data to fit an accurate estimator, especially the deep neural networks. Only net load data is used to disaggregate unknown PV generation based on the inherent temporal and spatial correlations [16], [19]. However, the distribution grids may have unobservable areas on edge, for which we have no access to complete smart meter data [9].

While the primitive information in distribution grids is limited, it can be inferred from extra public data sources that are easy to access, e.g., satellite images of PVs. Previous efforts lead to several approaches to identifying PV locations from Geographic information system (GIS), which is an image

References	Target	Data Sources	Prior Requirement	Model-Free Methods
[14]	PV Detection & Generation Estimation	Smart Meter Cloud Cover Index	Precise Predefined Hyperparameters	Change-Point Detection Algorithm
[15,16]	PV Detection & Generation Estimation	Smart Meter	Access to All-Customer Net Load Data	SVM Classification & LSTM Regression
[17]	PV Generation Estimation	Smart Meter	Sufficient Labeled Data	K-Means & PCA Feature Selection, and Linear Regression
[18]	PV Generation Estimation	Smart Meter Weather Conditions	Available PV Locations & PV Physical Model	Integrated Learning and Physics in SunDance
[19]	PV Generation Estimation	Smart Meter	Access to All-Customer Net Load Data	Gaussian Process Regression
[20]	PV Size, Tilt, and Azimuth Estimation	Smart Meter	Sufficient Labeled Data	Deep Neural Networks

Fig. 1: A summary of model-free PV detection and generation estimation works using smart meter data.

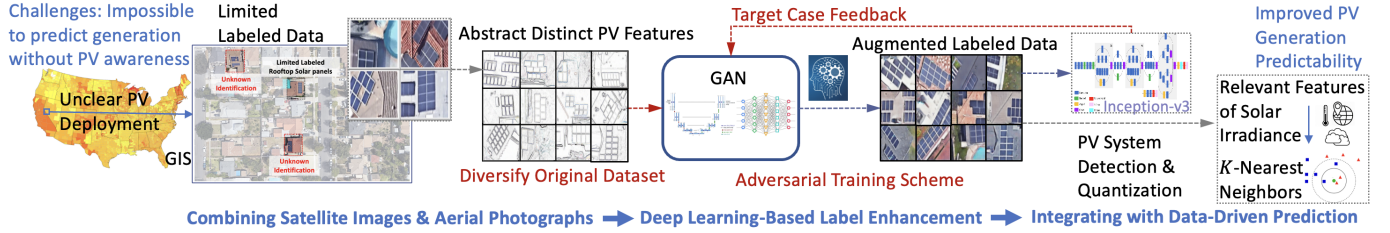


Fig. 2: An overview of the proposed approaches.

classification task [24]–[28]. However, the informative image data has complex structures to provide precise inference, and the insufficient labeled data in practice makes the classification method difficult for implementation.

For example, [24] and [25] train and test using similar PV system images without considering significant variations in solar panel positions, textures, numbers, and backgrounds. Though the rooftop solar detection tool based on convolutional neural networks (CNNs) claims better feature extraction for classification [26], [27], they rely on abundant hand-annotated data sets and are unclear about how to obtain the PV coordinates effectively [26], [27], [29]. Hence, these approaches are limited to solving the problem of locating and quantifying the distributed PVs in raw satellite images on their own.

Furthermore, a precise PV generation estimate helps maximize the information gain for DSOs. In the literature, [30] considers solar irradiation for an extended time period at a particular region to estimate generation. However, it is limited to addressing the concern of feature limitation. Also, [31] has introduced a support vector machine (SVM) as a base learner together with a meta learner, based on the K -means algorithm to cluster the training set for predicting short-term solar power generation. Such a methodology requires labeled historical PV data of the same solar power system, which is the same assumption of deep learning models such as probabilistic neural networks (PNNs) [32]. Therefore, previous methods find it difficult to extrapolate information, providing a solution when some PVs do not come with complete historical recordings.

To address the problems, we propose to utilize different data sources of publicly available information together with utility-owned measurements to enhance PV localization and genera-

tion forecast. Specifically, we use available GIS and satellite image data for PV detection and quantification. While existing works have also utilized such data sources, they have not considered the practical problem of labeled data deficiency as the first step, not to mention generation estimation. Therefore, we enhance PV localization and quantification by designing weak-supervised generative adversarial networks (GANs).

The proposed model not only generates diversified labeled data to address data deficiency but also embeds PV characteristics during generation to enable distinct PV image augmentation. The PV images are augmented for more accurate PV detection. Instead of separating two tasks, we integrate them into one model as a feedback loop, which makes improvements simultaneously for both data augmentation and PV detection during training. On the input side, we provide backbone structures as informative inputs to the generative model. The proposed model not only generates diversified labeled data to address data deficiency but also embeds PV characteristics during generation to enable distinct PV image augmentation. The PV images are augmented for more accurate PV detection in the downstream task. Instead of separating two tasks, we integrate them into one model as a feedback loop, which makes improvements simultaneously for both data augmentation and PV detection during training. Based on the detected PV information, we adopt the K -nearest neighbors method to estimate the best possible generation considering the relevant factors of solar irradiance. Fig. 2 shows an overview of the proposed method for generation predictability enhancement.

Our contribution lies in designing a constructive GAN to expand both the volume and diversity of labeled datasets for PV detection and enhancing generation estimation by utilizing effective features and neighboring generation patterns. Specifi-

cally, the learning performance of detection and quantification is boosted via specific knowledge embedding:

- 1) using diversified backbone structures (grey-scale images with different PV quantities and locations) as additional informative inputs;
- 2) integrating the evaluation of solar panel classification into the data augmentation process as guidance;
- 3) adopting both content and style losses to train the GAN model specifically for rooftop solar panel datasets.

II. FORMULATE LEARNING PROBLEM TO PREDICT DISTRIBUTED SOLAR POWER FROM (DIVERSIFIED) DATA

To predict widely distributed solar power, we need to accomplish three steps: 1) detect the PV existence with respect to locations, 2) quantify the amounts, and 3) estimate the output generation. While we have no access to direct information, it is inherently contained/held in diversified data sources. We fully utilize these data sources to infer the power output of distributed PVs in this paper. The task of latent information discovery is to solve a comprehensive/structural machine learning problem. Specifically, we start with the publicly available data sources (i.e., Google Earth) to sample raw satellite images (resolution of 4800×2987) with coordinates in the geographic information system (GIS). Since the raw image covers a large geographical area, we segment each one into $M = 400$ pieces to zoom into rooftops. Fig. 3 shows one



Fig. 3: A real image acquired from Google Earth and sliced.

sample image, where a few locations of the raw images have solar panels. With $M = 400$, less than 2% of segmentation is labeled as 1.

For segment $i \in \{1, 2, \dots, M\}$, the label $q_i \in \{0, 1, 2, \dots, k\}$ indicates the amount of solar panels, which is used in a multi-class classification for PV quantification. For a simplified PV detection task, the label is $q_i^e \in \{0, 1\}$ to indicate the non-existence and existence of solar panels in an image. Moreover, y_t , where $t = 1, \dots, T$ time points, is the generation of the corresponding solar power system based on the generation g_t^K of K nearest neighbor PVs. For generation estimation, other available data include the

temperature, humidity, and cloud cover with respect to the coordinates, which are the same location information used to sample satellite images. Therefore, the problem setup is as below.

- **Problem:** location detection, quantity assessment, and power generation estimation of rooftop PV systems;
- **Given:** 1) raw image $\mathbf{x}^{raw}$ from Google Earth and available labels q_i , 2) the known PV generation time series g_t^K of solar panel systems covered in the $\mathbf{x}^{raw}$ solar panel image, and 3) time-series weather data f_t^K of locations covered in the $\mathbf{x}^{raw}$ image;
- **Find:** for a new single segmented image x_i , 1) existence q_i^e of solar panels, 2) quantity q_i of solar modules, and 3) power generation time series $y_{i,t}$.

This paper aims to estimate distributed PV generations based on accurate PV location detection and quantity assessment. A robust image detector requires a large quantity of diversified data to train, while the known installation locations from the utility are limited to less than 2% of the total segments. To enlarge the labeled image data, an intuitive way is to use augmentation techniques (e.g., flip, rotate, extract patches, and transform color spaces) to operate on obvious data invariants [33]. For the solar panel case, we can rotate slices of images to contain different orientation angles of PV installations. This data augmentation is realistic, but there are no new instances created for data variability/diversity. The detection model may easily cause overfitting and perform poorly in unseen data. For a diversified image augmentation, deep learning-based generative models, such as generative adversarial networks (GANs), are popular at generating similar but varied samples as compared to the existing instances. GANs gain benefits from rich feature extraction of neural networks and the adversarial-training scheme; however, the lack of learning guidance causes the model to collapse easily, which creates bad samples like clustered solar panels pixels and object mixtures. Moreover, the labeled data is augmented with the goal of improving the training of a classification model (PV detection and quantification). The separation of the two steps can lead to propagation errors. To address these concerns, we aim to integrate both tasks into one compact learning model and train the entire model with mutual benefits to achieve high accuracy of PV classification against unbalanced data.

III. PROPOSED VARIATIONAL GANS WITH WEAK SUPERVISION FOR PV DETECTION AND QUANTIFICATION

To address the concerns of insufficient labeled data, we propose in this section an enhanced solar image classification based on weak supervision over GANs, as shown in the middle of Fig. 2.

A. Recall Basic GANs to Augment Image Samples

As mentioned in Section II, the goal of augmentation is to generate various solar images that mimic the original data in feature patterns as well as provide diversity.

To achieve this, we train a generator G as a feed-forward neural network parametrized by θ to produce new data x^{aug} ,

e.g., a rooftop solar panel image. x^{aug} is a random variable, whose complex distribution is expected to be learned from the distribution of real data samples x^{real} . In basic generative adversarial networks (GANs), the generator captures the mapping from a random variable z to image x^{aug} , where z is usually sampled from Gaussian noise distribution with the same size of x^{real} . To enforce the similar feature pattern in augmented images, the generative model G is trained discriminatively against another neural network, a discriminator D . The discriminator is parameterized by ϕ to score the outputs (comparing x^{aug} with x^{real}), computing the probability that one image comes from the real dataset. It aims to assign a high score to a real image x^{real} while assigning low scores to the generated image x^{aug} . Mathematically, the discriminator is realized as a classifier to maximize the binary cross-entropy loss while increasing the distinguishability between real and fake data. In contrast, the goal of generator G is to produce outputs that achieve high scores from the discriminator D , satisfying the constraints imposed by D in the process. Therefore, G minimizes the loss. Thus, for training the GANs, the objective takes the expectation of random variables as

$$L_{GAN} = \mathbb{E}_{x^{real}} [\log D(x^{real})] + \mathbb{E}_z [\log(1 - D(G(z)))], \quad (1)$$

and the training optimizes over $\min_{\theta} \max_{\phi} L_{GAN}$. Such an adversarial-training scheme can augment images on a larger invariant space, which is implicit for the data space.

GAN has become popular in solving power system problems because it adopts an adversarial strategy in generative learning to augment the diversity of labeled data. [34] firstly leverage the characteristic to generate scenarios of renewable outputs, which can mimic diverse conditions and uncertainties to produce more renewable data. Similarly, the GAN model is used to estimate the unknown power injection at unobservable loads based on the available historical measurements [35], [36]. While the generation from Gaussian noises is insufficient to cover the needs of specific data patterns, conditional inputs are used to restrict the generated data to a particular class like weather conditions of high wind, real-time system configurations of topology/admittances, and electricity market data [34], [35], [37], [38].

B. Embed PV Characteristics in Image Augmentation to Enhance Detection

The basic GANs generate new instances from noise inputs z to provide flexibility in the output. However, it easily causes insufficient learning of complex image distribution, especially when the labeled PV image data is limited for discriminative learning. The direct application of GANs creates bad samples like clustered solar panels pixels and object mixtures in the background. The generation process needs to be materialized with more explicit information. Moreover, unlike the assumption of sufficient training data in other GAN implementations, our task has the precondition of limited labeled solar panel images. And, the expected outputs should have distinct solar panels located on the roof, trees, pools, and other objects in the background.

Therefore, we adopt the conditional settings of GANs to consider extra information together with Gaussian noises.

Specifically, y is an embedding variable to condition the generative model on the external information. In our case, the solar panels naturally have the characteristic shape to be distinguished from other objects so we use the grey-scale outlines for y . The joint inputs of y and noise z contribute to replicating real PV images with both “backbone” representation and flexible background, as presented in the left-hand side Fig. 4. Similar to the basic GANs, the generator improves during training to fool the discriminator for a relatively high score compared to the real images. Mathematically, the generators in conditional GANs learn the mapping $\{y, z\} \rightarrow x^{aug}$, where y is the given information, z is the random noise, and x^{aug} is the output image. Meanwhile, discriminator D is trained to distinguish between the real and generated images, for which Fig. 4 illustrates the adversarial training scheme in the middle part. D is also conditional in such settings. It is fed with the concatenated image and the corresponding “backbone” information y . In the implementation, we have compared the conditional and unconditional settings of D , and it appears that conditional D better leads the augmented image x^{aug} to follow the PV characteristics. Thus, we have modified the objective for training GAN in (1) with consideration of extra information

$$L_{Aug} = \mathbb{E}_{x^{real}, y} \left[\log D(x^{real}, y) \right] + \mathbb{E}_{y, z} \left[\log \left(1 - D(y, G(y, z)) \right) \right]. \quad (2)$$

Despite the benefit, the input of grey-scale outlines has reduced the information on PV characteristics, which mainly embed the shape and position. To compensate for the remaining information that is essential for PV classification, we consider regularizing the generation process with limited but available labeled data. Although GANs do not use an explicit loss function, adding a traditional loss could benefit image generation. Motivated by [39], we consider the content loss and style loss to retain consistent PV characteristics on the pixel level and in the feature space. In the PV image generation task, the content loss is the mean square error between the generated image and the real image at the pixel level. Minimizing the content reconstruction loss recovers the detailed pixel information when the pair of grey-scale outline and ground-truth images are available. Notably, using the content loss will only have blurring and deterministic output images, which makes the conditional inputs of GANs in (2) necessary for our task. Moreover, while the labeled PV image data is limited, we use style loss minimization to guarantee style consistency from feature space among different images. Thus, the loss of image consistency of PV characteristics is

$$L_{Im} = \alpha L_{Content} + \beta L_{Style} \\ = \alpha \mathbb{E}_{x^{real}, y, z} [\|G(y, z) - x^{real}\|] + \beta \mathbb{E}_{x^{real}, y, z} [\|A^{G(y, z)} - A^{x^{real}}\|]. \quad (3)$$

Similar to the content loss, the style loss is the error between the features correlations expressed by Gram matrices $A = \sum_{k=1}^K F_{ik} F_{jk}$ (F is the feature map of the image). The hyperparameters α and β are weights to determine the emphasis of the two losses.

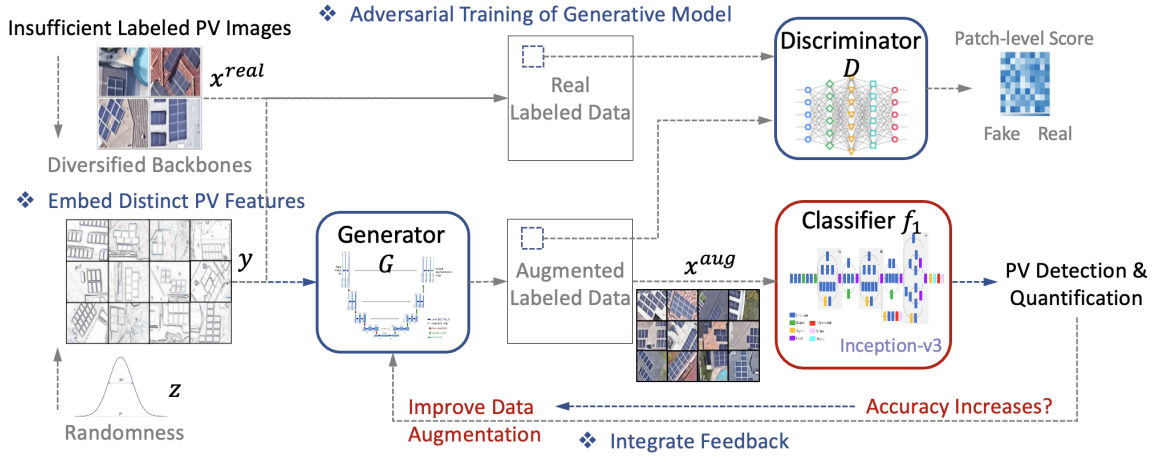


Fig. 4: Block diagram of the proposed variational GANs with weak supervision for PV detection and quantification.

C. Integrate Classification Feedback into PV Image Augmentation for Mutual Benefits

The generator is enhanced by embedding prior knowledge of PV characteristics on both inputs and loss functions. Besides, GANs use the discriminator to judge the quality of generated data by assigning scores. Although the discriminative learning scheme benefits, the DNN-based discriminator can be too general to control the generation process as expected. In our case, the specific goal of image generation/augmentation is to improve the accuracy of PV detection and quantification. GANs aim to enlarge the data to train the classifier for rooftop solar panels. Usually, data augmentation and classification are separate steps, where the former step cannot receive performance feedback from the latter task. It is difficult to make an improvement based on causal reasoning, and propagation errors between steps may exist. To solve the problems, we propose to integrate the two steps in a loop, which guides the augmentation towards improving the classifier f_1 . As Fig. 4 shows in the bottom half, we feed the augmented images from the generator into the f_1 and include the classification loss during training,

$$L_{Class} = -\mathbb{E}[\log P_{\eta}(x^{aug})], \quad (4)$$

where η parameterizes the classifier network. P is the probability of x to be class 1, which means the solar panel exists in the augmented image. In this way, the evaluation metric constrains the generation process. Since the classifier f_1 is integrated into the GAN structure, we recognize it as the other discriminator to distinguish augmented images that can improve PV classification.

In short, the final optimization is $\min_{\theta, \eta} \max_{\phi} L_{Aug} + L_{Im} + \lambda L_{Class}$. λ is the hyperparameter (defaulted as 10 in this paper) to specify the weight of the classification loss compared to the generator loss L_{Aug} . In this way, the labeled satellite images are effectively augmented to improve detecting PV systems.

Moreover, since we utilize task feedback as weak supervision, the PV classifier is trained via minimizing L_{Class} . During the training process, the generator and the classifier

are updated in turn for multiple iterations and converge. After learning, we simultaneously finish dataset augmentation and classification, so there is no extra step to train an additional model for the solar panel detection task.

D. Specify Conditional Inputs to Represent Distinct PV Features

The self-enhancement and feedback control discussed above provide weak supervision for generating images of rooftop solar panels. Not only increasing the volume of labeled data is beneficial, but also diversity is essential. For example, an effective classifier training expects images with different numbers of solar panels and various background objects. Although the available training dataset never has such instances, we introduce the randomness function to form the backbone, i.e., grey-scale outline as shown in Fig. 4. To enable it, we first apply different abstractions to the real labeled images, such as Sobel filter, semantic segmentor, landmark extractor, and color-specific filter [40], [41], after which we crop the basic backbone of the solar panel target. We then define t and s as the center and size (amount) of the single target, and r is the rotation degree if needed. The backbones of the diversified samples are constructed by different values of t , s , and r . The reconstructed backbone images are fed into the proposed variational GANs.

E. Proposed GAN Architectures and Optimization for PV Classification

The previous designs involve different deep learning functions to enhance PV detection and quantification. In this section, we specify the model architectures of each function and illustrate the training setups.

1) *Generator G for Image Generation*: To embed PV position and shape, we input conditional “backbones” to the generator. The desired output should not only retain the localized information from the grey-scale images but also reach high precision in context for classification. To enable such a transformation between high-dimensional inputs and outputs, we adopt the U-shaped architecture [42]. Its upsampling layers

balance the accurate object localization and context usage, which is capable of generating distinct features of solar panels.

2) *Discriminator D with Markovian Consistency*: As the consistencies in both pixel level and feature space are important for solar panel images, the discriminator follows the same rule. Different from the regular classifier, the whole image (real or generated) is scored in the unit of patch identically and independently, which is an $N \times N$ ($N \in \{1, 16, 70\}$) [43] square of the image. A smaller N focuses on the pixel level to benefit color representation, and a larger value sharpens spatial statistics across features. For example, $N = 1$ is the special pixel-level assessment, but the generated images cannot bring greater color diversity to our task. We find $N = 70$ works best to generate distinct objects in solar panel images.

3) *Classifier f_1 for PV Quantification*: The classifier is the last and most important step to detect and quantify solar panels in images. Meanwhile, it is expected to weakly supervise the generation process by feedback evaluation. The model Inception-v3 achieves state-of-the-art performance in image classification, whereas the feature extraction layers require a large and diversified dataset to train. Therefore, we pre-train the Inception-v3 model with 1.28 million images containing 1,000 different classes in the ImageNet and achieve 93.3% accuracy [44], [45]. We reuse the feature extraction layers and retrain the last layer of decision-making with our generated solar panel dataset for evaluation.

4) *Configuration*: During training, we use the Adam optimizer with a learning rate of 0.0002, and momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$ to train for 200 epochs for each experiment. For each epoch, the batch size is chosen as 5 due to the device limit. All the experiments are completed with a computer equipped with Inter(R) Core(TM) i7-9700k CPU and Nvidia GeForce RTX 3080Ti GPU.

IV. ESTIMATE GENERATION OUTPUT VIA CLOSE PROXIMITY OF PV SYSTEMS

Knowing the quantity of solar panels at a specific location helps gain insight into predicting generation capability, but its effectiveness needs to be further enhanced by historical PV generation profiles. Thus, to estimate solar generation output, we integrate the proposed PV detection and quantification with the data-driven solar irradiance forecasting, as shown in the right-hand side of Fig. 2. Specifically, based on historical observations, the PV generation patterns of neighbor installations are similar because of the similar numerical weather conditions. In residential areas, the neighboring houses are usually covered in one basic spatial unit to have the same feature values of numerical weather predictions. Therefore, this section describes the feature selection and generation estimation based on neighbor information.

A. Select Relevant Features Based on Weather Conditions

There are many different features available for making generation estimations. For example, the geographical features of longitude, latitude, altitude, weather conditions of temperature, pressure, humidity, cloudiness, and the quantities of modules. To increase the information gain and select the most relevant

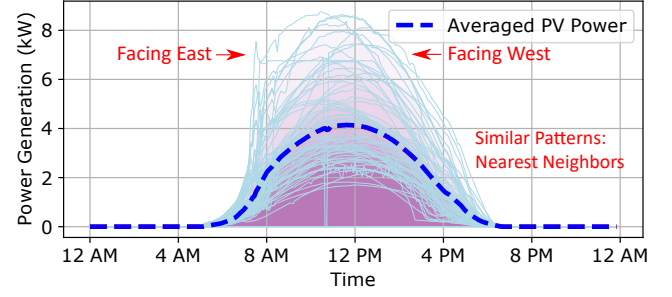


Fig. 5: Plot of raw data from the industrial partner SunPower Inc. shows homogeneous curves of power generation for three different ZIP code areas on a single day.

features, we followed a filter method for feature selection based on the information gain of minimum descriptive length (MDL) [46]. Let $a_{\cdot, \cdot}$ denote the number of training samples. $a_{i, \cdot}$ is the number of training samples from class P_i , and $a_{\cdot, j}$ gives the number of samples with the j^{th} value of the provided feature. Therefore, $a_{i, j}$ is the number of instances from class P_i and has the j^{th} value of the provided feature. If we have P classes, the information gain MDL is defined as a logarithm of all combinations of class labels

$$\text{MDL} = \frac{1}{a_{\cdot, \cdot}} \left(\binom{a_{\cdot, \cdot}}{a_{1, \cdot}, \dots, a_{P, \cdot}} - \sum_j \log \binom{a_{\cdot, j}}{a_{1, j}, \dots, a_{P, j}} \right) + \log \binom{a_{\cdot, \cdot} + P - 1}{P - 1} - \sum_j \log \binom{a_{\cdot, j} + P - 1}{P - 1} \quad (5)$$

A high score of information gain on geographical space shows that neighboring regions have similar PV generation patterns. The quantity of the modules is already available from previous learning results and serves as one of the most important variables, in that more solar panel units generate more power.

Moreover, to further boost the information gain for generation estimation, we consider the thermal space. The temperature is correlated with the efficiency of solar cells [47], which is also the efficiency of power generation. Based on the feature set, we learn to predict the power generation time series of the unknown solar power system.

B. Apply Nearest Neighbor Approach

With a set of carefully selected relevant features, we aim to fully utilize the latent correlation to make a precise estimation of PV generation. Fig. 5 shows a clear correlation among generations of the “neighboring” PV systems that locate in different ZIP (Zone Improvement Plan) code areas. The Euclidean distance is a promising measure for the distance between any two samples in feature space, not limited to only geographical distance [48]. With identified PVs in the previous learning process, the coordinates are known. Let q_i , x_i^{latitude} and $x_i^{\text{longitude}}$, and x_i^{weather} , $i = 1, \dots, N$, form the features of S_i^F that are selected from IV-A. Now, the

distances of a new test sample S_{test}^F comprising of features q_{test} , $S_{test}^{latitude}$, $S_{test}^{longitude}$, and $x_{test}^{weather}$. The data of different features are standardized to have a value range between 0.0 and 1.0 in preprocessing. $S_{test}^{altitude}$ and $S_{test}^{temperature}$ with K number of training samples can be calculated in a Euclidean manner. $d^K(S_K^F, S_{test}^F) = \sqrt{\sum_{f=1}^F (S_K^f - S_{test}^f)^2}$, where $K = 1 \cdots K^*$ and d^K is the vector of feature distances. In the feature space, the closest K points to S_{test}^F are used for training purposes. Their respective distances are used as the weights d_K and the PV generation time series of K nearest neighbors are represented as g_t^K . For $t = 1, \dots, T$ time points, the PV generation of the unknown solar power system is given by y_t . Thus, the optimization problem becomes

$$\min_{d,g} \sum_{i=1}^N \sum_{K=1}^{K^*} \sum_{t=1}^T |y_t - d^K g_t^K|^2. \quad (6)$$

This procedure gives an estimate of power generation for the unknown rooftop solar power system. We select the hyperparameter of K through two different error metrics, namely mean square errors (MSE) and mean absolute percentage errors (MAPE) respectively. MSE is calculated using the predicted power generation time series y_t and real power generation $y_{t,real}$, where $t = 1, \dots, T$ are the sequential time slots $MSE = \frac{1}{T} \sum_{t=1}^T (y_{t,real} - y_t)^2$. The variance of the distribution can be captured in mean absolute percentage error (MAPE) $MAPE = \frac{1}{T} \sum_{t=1}^T \frac{|y_{t,real} - y_t|}{y_{t,real}} \times 100\%$.

V. NUMERICAL TESTS

To evaluate the proposed methods on PV detection and generation estimation, we conduct experiments on realistic test cases in the following.

A. Data Pre-Processing and Tool Selection for Validation

First, we aim to enlarge the labeled solar panel image data so that we can train a classifier to detect the rooftop solar panel location from the geographic information system (GIS). We collect the original data from SunPower Inc. and the local utility, which contains the known installation locations of rooftop solar panels in GIS. With the longitudes and latitudes, the satellite images are sampled with a fixed resolution of 4800×2987 on Google Earth (fixed zoom level). It is free and has several benefits when compared to its commonly used rivals like SPOT5 or DigitalGlobe satellite imagery. [49] illustrates the high-resolution imagery of the Google Earth archive has robust positional accuracy and plays a vital role in solar panel image recognition and quantification. The original positive sample number is only $1017 + 2813 = 3830$, while related detection work used tens/hundreds of times of positive samples for training. The classification performance is greatly affected since the limited image data cannot cover the scenarios that exist in reality. Due to the high unbalance in the original image dataset, we select partial images without solar panels to be negative samples and get 15000 as a total. Moreover, we use the geographic coordinates to integrate a time series of power generation data ($612 + 239$ data sequences). Typically, the recorded time series of PV data

is with a time resolution of 15 minutes for a whole year. So, to align the data format we chose the time resolution of weather conditions to be 15 minutes as well. We aim to show robustness by validating the method in two different states in the U.S., 1) Tempe, Arizona, and 2) Santa Ana, California.

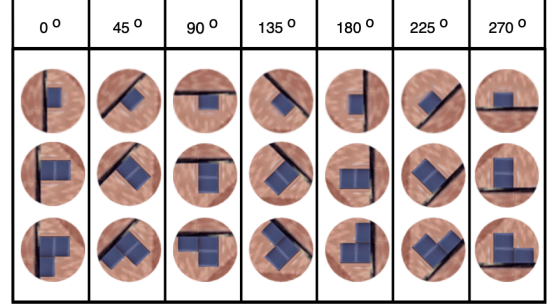


Fig. 6: The classic augmentation of solar panel image via basic operation of circular rotation.

B. Enriched Labeled Data with Physics Embedding Improve Detection Accuracy

In the following, we validate the accuracy of PV detection and quantification of the proposed weakly supervised GANs.

1) *Effective Augmentation*: To solve the data deficiency problem, we first consider standard augmentation techniques with respect to real conditions that residential houses can face from different directions, and the house styles can be quite similar within each community. Therefore, by adequately rotating images via the rotation matrix $R = [\cos \theta, -\sin \theta; \sin \theta, \cos \theta]$, $\theta \in [0, 2\pi]$, we increase the labeled dataset. Since direct rotation changes the horizontal structure of the image, we pre-process the data to crop them circularly, using the circular segmentation-based approach as in [50], where the shape will be consistent no matter how it is rotated (Fig. 6). In such augmentation, the information around the edges and vertices is compromised, and the performance suffers. Meanwhile, the improvement of the classifier requires sufficient and various positive samples, e.g., images with solar panels in different scenarios. Therefore, we expand the labeled dataset by the GANs.

We validate the proposed GAN with two types of metrics on the generated data quality. On one hand, Fréchet Inception Distance (FID) [51] measures the closeness of extracted features between generated data and real data for image generation tasks. Namely, the lower FID score reveals more similarity, which is a better quality of augmented images. On the other hand, the image dataset is enlarged to boost classification performance so that we use the test performance of the solar panel classification to assess the augmentation quality. The output labels of the classifier f_1 are for the quantification task, and we compute the accuracy by comparing the f_1 outputs with ground truth labels, which is:

$$\text{Quantification Accuracy} = \frac{\text{Amount of Samples}\{\hat{q}_i = q_i\}}{\text{Amount of Total Samples}}.$$

We further compute the commonly used metrics to evaluate the detection performance, which are

$$\text{Detection Precision} = \frac{TP}{TP + FP}, \quad (7)$$

$$\text{Detection Recall} = \frac{TP}{TP + FN}, \quad (8)$$

$$\text{Overall Accuracy} = \frac{TP + TN}{\text{Amount of Total Samples}}. \quad (9)$$

The “true” and “false” represent whether or not the classification results are the same as labels, while “positive” and “negative” mean with and without solar panels in the image as two classes. Namely, “true positive (TP)” represents that solar panels exist and are correctly detected. Similarly, “false positive (FP)” means the positive (wrong) estimation of negative samples. For our task, we aim to evaluate the model’s capability of correct solar panel detection. First, we expect to know what proportion of the detected solar panels does actually exist. Thus, the detection precision in (7) measures the ratio of correct solar panel detection among all the positive output (with solar panel) samples. It is used to evaluate the quality of augmented positive examples, as shown in Table I. Meanwhile, we are interested in what proportion of the images with actually existing solar panels are detected. This is computed by the detection recall in (8). Moreover, we include 1) the overall classification accuracy to represent the correct classification for both positive and negative samples, and 2) the quantification accuracy to represent the correct identification of the amount of existing solar panels.

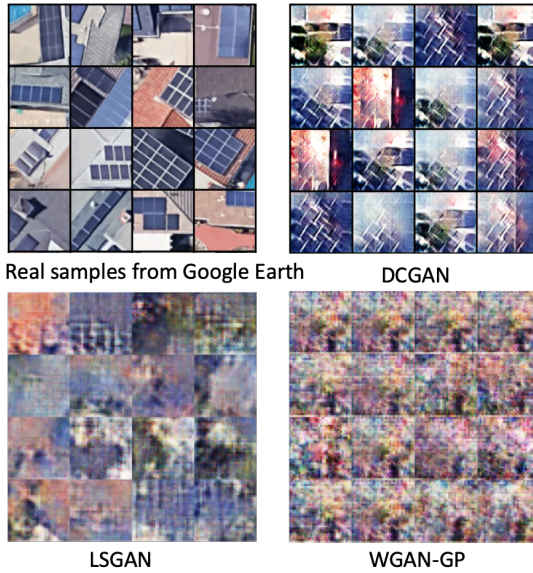


Fig. 7: Generated results after training the popular architectures of GANs. There appears to be visually under-fitting and missing training direction via repeated noise textures across multiple samples.

First, we tried popular architectures, such as DCGAN, LSGAN, and WGAN-GP [52]–[54]. However, the learning of mapping from simple noises to target images poorly follows our expected direction of distinct solar panels in images, as

shown in Fig. 7. Thus, the proposed model improves upon two aspects: providing informative inputs and embedding feedback evaluation.

2) *Informative Inputs for Self-Enhancement*: In this task, dependent information y is the grey-scale image that covers the number of solar panels and their positions in an image. We adopt the Pix2Pix model [55] as a basis for its superior image translation. The abstraction of grey-scale images is an edge detection task, for which we select the Sobel filter among different filters [40]. It convolves the image with a small, separable, and integer-valued filter in the horizontal and vertical directions and is relatively efficient in computations. The Sobel filter enhances the edges of objects in grey-scale images by providing differentiating (which gives the edge response) and smoothing (which reduces noise) concurrently.

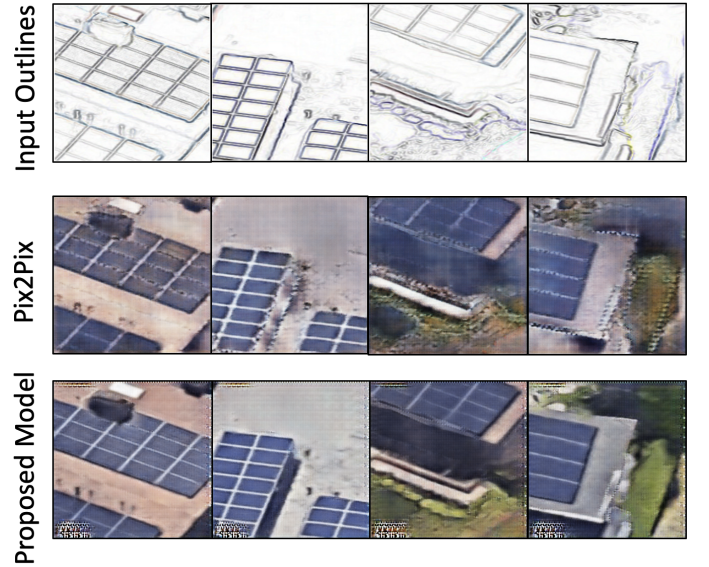


Fig. 8: Given conditional inputs of “backbones” (top), compare the generated rooftop solar panel images from Pix2Pix (middle) and our proposed model (bottom). Without specific guidance, the solar panels easily mix with background objects.

During training, we found that using the base model (Pix2Pix) only makes the generator hard to converge. The discriminator converges fast and cannot further improve the generation, because our solar panel data is complex with multiple objects. Such data complexity can also be observed from a high base FID value (131.07 ± 3.47), for which we randomly separate the real solar panel dataset into two groups and measure the similarity in between them. Therefore, we increase the depth of discriminator D to balance with the generator capability and use more task-specific losses as proposed, leading to the complete model in Table I. The solar images generated by the proposed complete model achieve a much lower FID score than the base data generation model (Pix2Pix). The visual comparison in Fig. 8 shows intuitive improvement. The images “translated” by Pix2Pix in the middle resemble the real ones from Google Earth, which are more comparable than those from standard GANs (Fig.

TABLE I: The quality of generated data after 200 epochs. A lower FID indicates better similarity and a higher accuracy means that the data augmentation better benefits the post classification task. To serve as a basis, the FID of two randomly separated real datasets is 131.07 ± 3.47 and the classifier trained with real data has the test accuracy of 0.59 ± 0.030 .

Model	Pix2Pix	Proposed Model			
		Complete Model	W/o $L_{Content}$	W/o L_{Style}	W/o L_{Class}
FID	171.50 ± 2.74	151.06 ± 0.73	174.56 ± 3.49	155.37 ± 1.34	156.18 ± 0.01
Detection Precision ($\times 100\%$)	0.81 ± 0.017	0.88 ± 0.054	0.62 ± 0.021	0.87 ± 0.013	0.82 ± 0.038
Detection Recall ($\times 100\%$)	0.83 ± 0.023	0.91 ± 0.005	0.60 ± 0.097	0.88 ± 0.034	0.81 ± 0.062
Overall Accuracy ($\times 100\%$)	0.84 ± 0.006	0.89 ± 0.068	0.68 ± 0.043	0.89 ± 0.045	0.84 ± 0.056
Quantification Accuracy ($\times 100\%$)	0.80 ± 0.092	0.87 ± 0.047	0.55 ± 0.089	0.87 ± 0.042	0.82 ± 0.101

7). Nevertheless, the generated solar panels often mix with the background objects, and the Pix2Pix model sometimes collapses around 100 epochs or even earlier. In contrast, our model with specific guidance can better recover the target solar panels in the image with clear edges and color, as shown in the bottom of Fig. 8. To demonstrate the contributions of each component, we conduct an ablation study. The first row of Table I shows that content loss plays an essential role in synthesizing images. We select $L_{Content}$ and L_{Style} , which both make a large difference in the FID score, as shown in Fig. 9. While the pixel-level loss can encourage color similarity of the image, it can also bring blurring results. The style loss better corrects the color of the target objects, leading to distinct solar panels. Moreover, the contribution to the following classification task is more essential for PV detection, which is revealed in the next rows of Table I, and we analyze it in the next section.

3) Feedback Evaluation to Constrain the Augmentation:

As the goal is to improve the classifier for solar panel detection, we channel the feedback from the state-of-art pre-trained Inception-v3. The images generated under such weak supervision are then fed into the classifier as positive samples, and the cross-entropy loss is fed back to the generator for control. The GAN and the inception-v3 classifier are trained simultaneously to augment positive samples. Specifically, the last four columns of Table I show the results of an ablation study of the proposed variational GAN model, where each column reveals the contribution of each design to the PV image augmentation and PV classification. Without any design, the performance deteriorates. To compare the third and the last columns of Table I, we observe that although integrating the classifier does not significantly improve FID on image similarity, it benefits the post-classification task with higher accuracy.

Except for the comparison in the ablation study, we explore the improvement of the classifier with different data availabilities to show the benefit of GAN-based augmentation. The following datasets are considered to train the classifier for PV detection and quantification: 1/2 of the original dataset with limited labels (1/2), the entire original dataset (original), the enlarged labeled dataset by basic functional image augmentation such as rotation and flip (basic Aug.), and different sizes of expanded labeled datasets from the proposed variational GAN (GAN Aug. $1\times$, $2\times$, and $3\times$), respectively.

Fig. 10 demonstrates the trends of loss during training. There are significant decreases in converged losses when training the classifier with the data augmented by the proposed

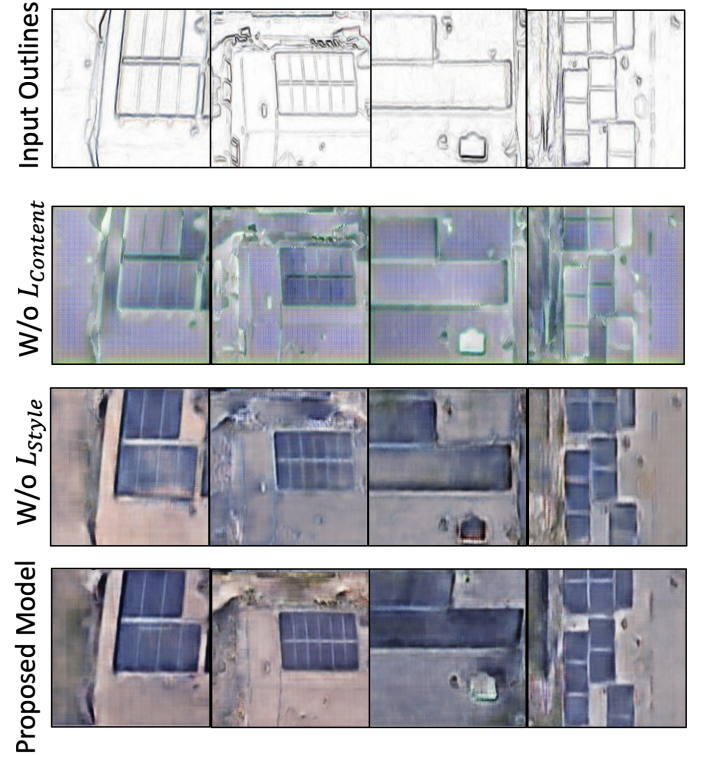


Fig. 9: Visual results of the ablation study based on the obvious differences in FID values (Table I). While the content loss incentivizes the color similarity of the whole image (pixel-level), the logical constraint focuses on the target objects more. Such weak supervision improves the generation quality, leading to distinct solar panels in the image.

GAN. Specifically, Fig. 10a - Fig. 10c present fast loss decreases before 50^{th} epoch, but the classification losses cannot be lower after then. In contrast, the training losses of Fig. 10d - Fig. 10f experience an increase and then go down. It is due to the simultaneous training of the classifier and the generator for image augmentation. In the first few epochs, the proposed GAN keeps updating parameters from random initialization, for which the generated labeled images are not perfect yet. Therefore, the classifier trained by these generated images has high losses. With the image generation improving, the loss of the classifier keeps decreasing until convergence. We observe that the losses close to convergence are lower and more stable. Moreover, we test the trained model on the

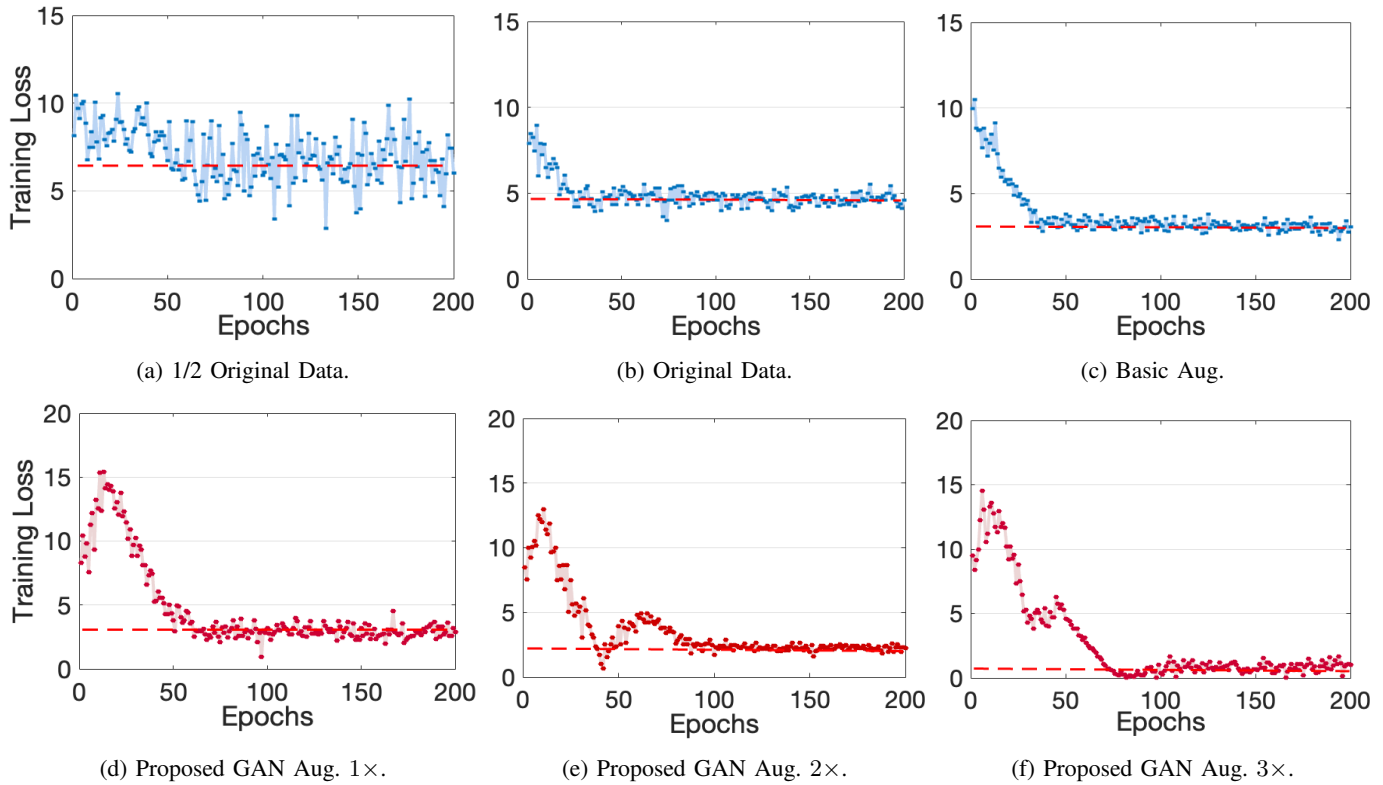


Fig. 10: Classification loss during training (epochs) using different training datasets with respect to augmentation methods: (a) 1/2 of the original dataset with limited labels, (b) the entire original dataset, (c) the enlarged labeled dataset by basic functional image augmentation, and (d) - (f) different sizes of expanded labeled datasets from the proposed variational GAN (GAN Aug. 1 $\times$, 2 $\times$, and 3 $\times$).

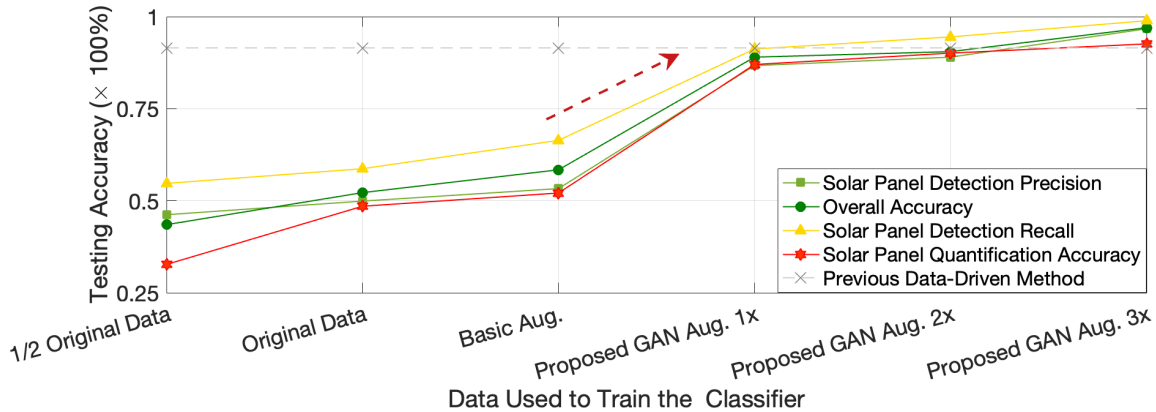


Fig. 11: The testing accuracy of PV detection and quantification when training the classifier f_1 using different training datasets with respect to augmentation methods.

randomly sampled images of the residential areas in California and Arizona and show the results in Fig. 11. Previous data-driven methods assume sufficient labeled data for training, so the accuracy is high when using only sufficient original datasets or basic augmented data. The adopted classifier f_1 , the Inception-v3 model, has state-of-the-art performance in image classification, which is a benchmark model. In our case, using only the original labeled dataset or dataset with basic augmentation, the accuracy is low for PV detection and quantification. We observe a noticeable increase in accuracy

from the barplot when using the proposed GAN method to augment labeled data. With the increasing amount and diversity of augmented data, the testing accuracy keeps going up and reaching comparable and even better performance than a previous data-driven method that assumes sufficient labeled data.

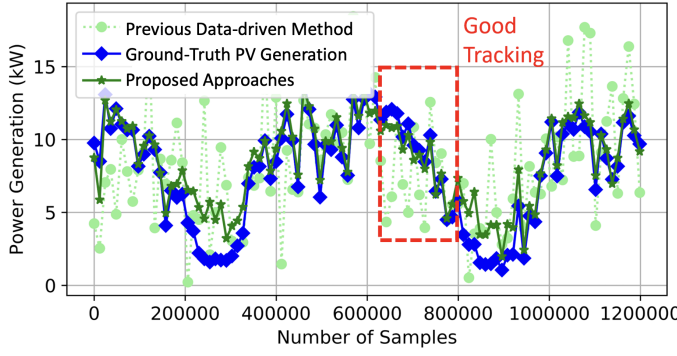


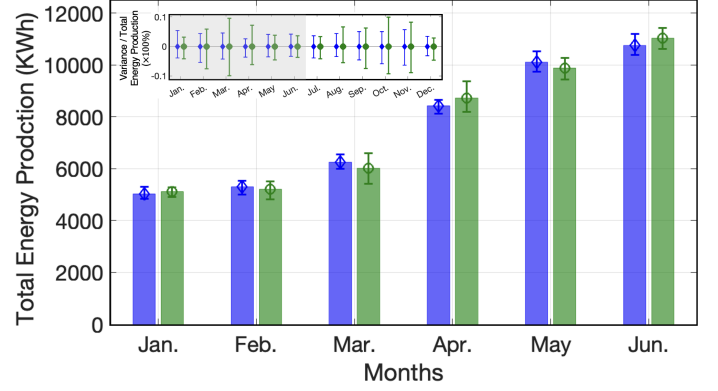
Fig. 12: Compare the real and the predicted downsampled PV generation of a rooftop solar power system.

C. Robust Generation Estimation via Flexible Features and K -NN

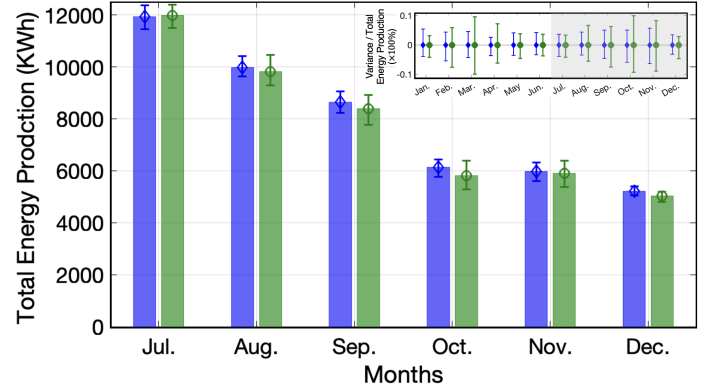
The validation to detect and quantify PV systems brings us to the next stage of generation estimation validation. There are many different features available for making estimations of PV generation. But, to maximize the information gain, we propose to select the most relevant features using a filter method-based approach. The result of that approach is a set of features that include geographical coordinates, weather conditions of temperature and cloudiness, and the quantity of modules. Furthermore, the regions differ in solar irradiance, climate, soiling profile, and terrain from one another [56]. The generation of such solar panels is provided by SunPower Inc. and is used as the training set for learning. We use the K -nearest neighbor (K -NN) method for learning by assigning the weights to the time-series data of the nearest points geographically. The only hyperparameter to be tuned here is K . We will present the performance changes with respect to K later.

Then, we apply the weighted K -NN regression to predict short-term solar power generation. Fig. 12 compares the predicted generation with the ground truth for an entire year. The previous data-driven method refers to the support vector regression (SVR) model, which has shown good performance in PV generation estimation [11], [12], [15]. Since the experiment setups are different, we use the SVR model but change the inputs to be the same as that of our case. Specifically, the SVR implementation does not have a GAN-based data augmentation to train a classifier for PV quantification as a preliminary step. To better visualize the comparison, both real and predicted generation data are downsampled. We plot the long-term accumulated generation based on months in Fig. 13. The results are expressed in a bar plot over 12 months in 2015, where the black error bars indicate the variance in generation estimation. From a general energy production point of view, the model performance is stable in the months when the weather is stable. For example, estimations of June, July, November, and December have higher accuracy than March, April, September, and October.

To select the best K data samples of neighbors for PV generation estimation, we compare the performances in Table II. Obviously, $K = 3$ is the optimal choice for both error



(a) The bar plot of generation from January to June.



(b) The bar plot of generation from July to December.

Fig. 13: Comparison of accumulated real and predicted monthly PV energy generations for the year 2015.

TABLE II: Compare errors of generation estimation with respect to different K values.

No. of Neighbors (K)	MSE	MAPE (%)
1	0.2422	24.14
2	0.2056	5.80
3	0.1528	5.21
4	0.1719	8.92
5	0.2062	19.85
6	0.2187	21.78

metrics. If the data samples of more or less nearest neighbors are considered, the estimation errors increase. Especially, if we select $K > 4$, the error is much higher, for which too much feature distance between the considered PV systems could increase the variance for estimation.

VI. CONCLUSION

To accommodate the limited data availability and timeliness of PV data in distributed power systems, we propose to systematically enhance PV localization and generation forecast using multiple data sources such as satellite imagery and numerical weather conditions. Specifically, we first design weakly supervised GANs for solar panel image augmentation.

Multiple aspects of GAN enhancement are designed to augment the images that can improve PV classification, including: 1) preparing backbone images as conditional inputs to embed PV characteristics, and 2) restricting the inexplicit learning process of the GAN model by specific losses. Moreover, we leverage the discriminative training mode of GAN to integrate PV detection and quantification into the augmentation loop. In this way, the performance of the targeted downstream classification task guides the image generation process. Thus, we obtain the detection results without further effort and combine them with historical neighboring measurements to estimate the PV generation. We validate the proposed approaches on areas of distribution grids that have wide PV coverage but limited prior information. The result shows that the proposed approaches can efficiently avoid model collapse in image generation, reach comparable classification performance with methods trained using sufficient data, and obtain accurate generation estimation.

REFERENCES

- [1] P. A. Basore and W. J. Cole, "Supply and demand constraints on future PV power in the USA," in *IEEE Photovoltaic Specialist Conference*, 2017, pp. 2163–2166.
- [2] S. Jerez, I. Tobin, R. Vautard, J. P. Montávez, J. M. López-Romero, F. Thais, B. Bartok, O. B. Christensen, A. Colette, M. Déqué *et al.*, "The impact of climate change on photovoltaic power generation in europe," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.
- [3] J. Yuan, Y. Weng, and C.-W. Tan, "Determining maximum hosting capacity for pv systems in distribution grids," *International Journal of Electrical Power & Energy Systems*, vol. 135, p. 107342, 2022.
- [4] S. Luo, Y. Weng, E. Cook, R. Trask, and E. Blasch, "Solar panel identification under limited labels."
- [5] J. Yuan and Y. Weng, "Support matrix regression for learning power flow in distribution grid with unobservability," *IEEE Transactions on Power Systems*, vol. 37, no. 2, pp. 1151–1161, 2022.
- [6] B. Saleem, Y. Weng, and F. Gonzales, "Association rule mining for localizing solar power in different distribution grid feeders," *IEEE Transactions on Smart Grid*, 2020.
- [7] J. Yuan and Y. Weng, "Physics interpretable shallow-deep neural networks for physical system identification with unobservability," in *2021 IEEE International Conference on Data Mining (ICDM)*, 2021, pp. 847–856.
- [8] A. Fleischhacker and *et al.*, "Sharing solar PV and energy storage in apartment buildings: resource allocation and pricing," *IEEE Trans. on Smart Grid*, vol. 10, no. 4, pp. 3963–3973, 2018.
- [9] J. Yuan, Q. Cui, Z. Ma, and Y. Weng, "Learning ev placement factors with social welfare and economic variation modeling," in *2019 North American Power Symposium (NAPS)*. IEEE, 2019, pp. 1–6.
- [10] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2015.
- [11] H. S. Jang, K. Y. Bae, H.-S. Park, and D. K. Sung, "Solar power prediction based on satellite images and support vector machine," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.
- [12] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Transactions on Industry Applications*, vol. 48, no. 3, pp. 1064–1069, 2012.
- [13] A. Catalina, C. M. Alafz, and J. R. Dorronsoro, "Combining numerical weather predictions and satellite data for pv energy nowcasting," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 3, pp. 1930–1937, 2020.
- [14] X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential pv installations," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2477–2485, 2016.
- [15] F. Wang, K. Li, X. Wang, L. Jiang, J. Ren, Z. Mi, M. Shafie-khah, and J. P. Catalão, "A distributed pv system capacity estimation approach based on support vector machine with customer net load curve features," *Energies*, vol. 11, no. 7, p. 1750, 2018.
- [16] J. Wang, W. Zheng, and Z. Li, "Detection and estimation of behind-the-meter photovoltaic generation based on smart meter data analytics," *The Electricity Journal*, vol. 35, no. 5, p. 107132, 2022.
- [17] H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2466–2476, 2016.
- [18] D. Chen and D. Irwin, "Sundance: Black-box behind-the-meter solar disaggregation," in *Proceedings of the eighth international conference on future energy systems*, 2017, pp. 45–55.
- [19] F. Bu, R. Cheng, and Z. Wang, "A two-layer approach for estimating behind-the-meter pv generation using smart meter data," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 885–896, 2023.
- [20] K. Mason, M. J. Reno, L. Blakely, S. Vejdani, and S. Grijalva, "A deep neural network approach for behind-the-meter residential pv size, tilt and azimuth estimation," *Solar Energy*, vol. 196, pp. 260–269, 2020.
- [21] A. Argüello, J. D. Lara, J. D. Rojas, and G. Valverde, "Impact of rooftop pv integration in distribution systems considering socioeconomic factors," *IEEE Systems Journal*, vol. 12, no. 4, pp. 3531–3542, 2017.
- [22] Z. Chi, L. Pandian, Z. Xueying, Z. Jie, Z. Wei, H. Jiajian, X. Yu, and X. Qi, "Research on the impacts of grid-connected distributed photovoltaic on load characteristics of regional power system," in *2017 International conference on green energy and applications (ICGEA)*. IEEE, 2017, pp. 95–99.
- [23] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states," *Joule*, vol. 2, no. 12, pp. 2605–2617, 2018.
- [24] J. M. Malof, Rui Hou, L. M. Collins, K. Bradbury, and R. Newell, "Automatic solar photovoltaic panel detection in satellite imagery," in *International Conference on Renewable Energy Research and Applications*, 2015, pp. 1428–1431.
- [25] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," *Applied Energy*, vol. 183, pp. 229 – 240, 2016.
- [26] V. Golovko and *et al.*, "Convolutional neural network based solar photovoltaic panel detection in satellite photos," in *IEEE International Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, vol. 1, 2017, pp. 14–19.
- [27] V. Golovko, A. Kroshchanka, S. Bezobrazov, A. Sachenko, M. Komar, and O. Novosad, "Development of solar panels detector," in *International Scientific-Practical Conference Problems of Infocommunications. Science and Technology*, 2018, pp. 761–764.
- [28] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," *Applied energy*, vol. 183, pp. 229–240, 2016.
- [29] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the United States," *Joule*, vol. 2, pp. 2605–2617, 2018.
- [30] M. Alanazi, A. Alanazi, and A. Khodaei, "Long-term solar generation forecasting," in *IEEE PES Transmission and Distribution Conference and Exposition*, 2016, pp. 1–5.
- [31] H. Zhou, Y. Zhang, L. Yang, and Q. Liu, "Short-term Photovoltaic power forecasting based on Stacking-SVM," in *International Conference on Information Technology in Medicine and Education*, 2018, pp. 994–998.
- [32] B. Ke, T. Ku, Y. Ke, C. Chuang, and H. Chen, "Sizing the battery energy storage system on a university campus with prediction of load and photovoltaic generation," *IEEE Transactions on Industry Applications*, vol. 52, no. 2, pp. 1136–1147, 2016.
- [33] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [34] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3265–3275, 2018.
- [35] M. Kamal, W. Li, D. Deka, and H. Mohsenian-Rad, "Physics-conditioned generative adversarial networks for state estimation in active power distribution systems with low observability," in *2022 International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA)*, 2022, pp. 1–6.
- [36] Y. Gu, Q. Chen, K. Liu, L. Xie, and C. Kang, "Gan-based model for residential load generation considering typical consumption patterns," in *2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2019, pp. 1–5.
- [37] G. Raman, C. J. O'Rourke, J. Lu, J. C.-H. Peng, and J. L. Kirtley, "Conditional generative adversarial networks for dynamic control-parameter selection in power systems," *IEEE Access*, vol. 10, pp. 11 236–11 247, 2022.

- [38] Z. Zhang and M. Wu, "Predicting real-time locational marginal prices: A gan-based approach," *IEEE Transactions on Power Systems*, vol. 37, no. 2, pp. 1286–1296, 2022.
- [39] L. A. Gatys, A. S. Ecker, M. Bethge, and C. Sep, "A neural algorithm of artistic style. arxiv 2015," *arXiv preprint arXiv:1508.06576*.
- [40] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014. [Online]. Available: <https://doi.org/10.7717/peerj.453>
- [41] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," pp. 1867–1874, 2014.
- [42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," pp. 234–241, 2015.
- [43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- [44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," 2009.
- [45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2016.
- [46] Q. Cui, K. El-Arroudi, and Y. Weng, "A feature selection method for high impedance fault detection," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1203–1215, 2019.
- [47] W. Yang, C. Lo, C. Wei, and W. Lour, "Cell-temperature determination in InGaP-(In)-GaAsGe triple-junction solar cells," *IEEE Electron Device Letters*, vol. 32, no. 10, pp. 1412–1414, 2011.
- [48] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean distance of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334–1339, 2005.
- [49] U. Paredes-Hernández and et al., "Horizontal positional accuracy of google earth's imagery over rural areas: A study case in Tamaulipas, Mexico," *Boletim de Ciências Geodésicas*, vol. 19, pp. 588–601, 2013.
- [50] P. Ayyalasomayajula, S. Grassi, and P. Farine, "Rotation, scale and translation invariant image retrieval method based on circular segmentation and color density," in *International Symposium on Image and Signal Processing and Analysis*, 2011, pp. 455–459.
- [51] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.
- [52] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.
- [53] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," 2016.
- [54] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
- [55] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016.
- [56] J. L. Kafka and M. A. Miller, "A climatology of solar irradiance and its controls across the United States: Implications for solar panel orientation," *Renewable Energy*, vol. 135, pp. 897 – 907, 2019.



Yang Weng (Senior Member, IEEE) He received the B.E. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, the M.Sc. degree in statistics from the University of Illinois at Chicago, Chicago, IL, USA, and the M.Sc. degree in machine learning of computer science and the M.E. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA. He joined Stanford University, Stanford, CA, USA, as the TomKat Fellow for sustainable energy. He is currently an Assistant Professor of electrical, computer, and energy engineering with Arizona State University, Tempe, AZ, USA. His research interest is in the interdisciplinary area of power systems, machine learning, and renewable integration. He received the CMU Dean's Graduate Fellowship in 2010, the Best Paper Award at the International Conference on Smart Grid Communication (SGC) in 2012, the First Ranking Paper of SGC in 2013, the Best Papers at the Power and Energy Society General Meeting in 2014, the ABB Fellowship in 2014, the Golden Best Paper Award at the International Conference on Probabilistic Methods Applied to Power Systems in 2016, and the Best Paper Award at IEEE Conference on Energy Internet and Energy System Integration in 2017, IEEE North American Power Symposium in 2019, and the IEEE Sustainable Power and Energy Conference in 2019.



Jingyi Yuan (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, and the M.Sc. degree in electrical engineering from Arizona State University (ASU). She is currently serving as a Research Assistant and pursuing a Ph.D. degree in electrical, computer and energy engineering at Arizona State University, Tempe, AZ, USA. In recognition of her extraordinary academic achievements, she won University Graduate Fellowship and Engineering Graduate Fel-

lowship. Her research interests are in the areas of machine learning and big data applications in power systems, explainable data-driven modeling of cyber-physical systems, and dynamic hosting capacity of photovoltaics. She won the Best Paper (Presentation) Award of the 51st IEEE North American Power Symposium in 2019.