10-2021

Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder

Meng SHEN

Huaizheng ZHANG

Yixin CAO
Singapore Management University, yxcao@smu.edu.sg

Fan YANG

Yonggang WEN

Citation

SHEN, Meng; ZHANG, Huaizheng; CAO, Yixin; YANG, Fan; and WEN, Yonggang. Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder. (2021). *Proceedings of the 29th ACM International Conference on Multimedia, Virtual Conference, 2021 October 20-24*. 2558-2566.
Available at: https://ink.library.smu.edu.sg/sis_research/7320

Missing Data Imputation for Solar Yield Prediction using Temporal Multi-Modal Variational Auto-Encoder

Meng Shen
Nanyang Technological University
meng005@e.ntu.edu.sg

Huaizheng Zhang
Nanyang Technological University
huaizhen001@e.ntu.edu.sg

Yixin Cao
Nanyang Technological University
yixin.cao@ntu.edu.sg

Fan Yang
Nanyang Technological University
yang.fan@ntu.edu.sg

Yonggang Wen
Nanyang Technological University
ygwen@ntu.edu.sg

ABSTRACT

The accurate and robust prediction of short-term solar power generation is significant for the management of modern smart grids, where solar power has become a major energy source due to its green and economical nature. However, the solar yield prediction can be difficult to conduct in the real world where hardware and network issues can make the sensors unreachable. Such data missing problem is so prevalent that it degrades the performance of deployed prediction models and even fails the model execution. In this paper, we propose a novel temporal multi-modal variational auto-encoder (TMMVAE) model, to enhance the robustness of short-term solar power yield prediction with missing data. It can impute the missing values in time-series sensor data, and reconstruct them by consolidating multi-modality data, which then facilitates more accurate solar power yield prediction. TMMVAE can be deployed efficiently with an end-to-end framework. The framework is verified at our real-world testbed on campus. The results of extensive experiments show that our proposed framework can significantly improve the imputation accuracy when the inference data is severely corrupted, and can hence dramatically improve the robustness of short-term solar energy yield forecasting.

CCS CONCEPTS

• **Computing methodologies** → *Learning latent representations*.

KEYWORDS

solar forecasting, multimodal learning, data imputation

ACM Reference Format:
Meng Shen, Huaizheng Zhang, Yixin Cao, Fan Yang, and Yonggang Wen. 2021. Missing Data Imputation for Solar Yield Prediction using Temporal Multi-Modal Variational Auto-Encoder. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475430

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475430

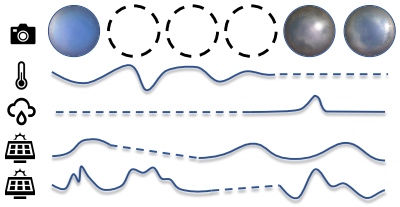


Figure 1: Data Missing and Mismatch among Multiple Modalities. Dashed lines represent the missing data.

1 INTRODUCTION

Solar power has emerged as one of the most attractive renewable energy sources for achieving carbon neutrality across the globe. According to the analytics provided by the IRENA (International Renewable Energy Agency) [7], solar power is popular in modern energy architecture due to three reasons: zero carbon emission, decreasing installation cost, and increasing capacity. Benefiting from the mature photo-voltaic technology, the electricity cost for utility-scale solar panels reached 0.068 USD per Kilowatt-hour(kWh) in 2019, with a remarkable 13% yearly reduction. Meanwhile, solar power took up nearly a quarter of the total installed renewable energy capacity in 2019 and the share continues to grow. However, the yield volatility of solar energy incurs substantial challenges for power grid operation and management. By nature, solar power generation is unstable and uncontrollable since it is mainly dominated by weather conditions. For instance, solar power yield can drop dramatically when rainstorms take place or cumulus clouds float above solar panels, blocking the solar radiation [18]. The growing penetration of unstable solar power supply in our modern electricity grids may invoke serious disequilibrium between energy supply and demand, by stressing the power grids, increasing their operational cost, and challenging their reliability.

To tackle the problem, solar yield prediction methods have been proposed as a mitigation measure for curbing yield volatility. These methods fall into two classes, short-term and long-term forecasts. In this paper, we mainly focus on the former (hours or minutes ahead prediction). Currently, the common methodologies in this category include: 1) building physical models of weather conditions and solar panels [2, 22], 2) employing ground-based sky cameras to capture hemispherical sky images [23, 25, 32], and 3) developing statistical models that analyze the historical trends of power generation and meteorological information [16, 17, 21]. These approaches could be

generally categorized as single-modality methods, analyzing either the graphic info of the sky or the numerical data of the environment. Recently, the emerging multi-modal learning scheme applied in other cross-domain modelings [3, 29, 30], such as language-vision, has been attempted in the solar yield prediction field [31] and achieved higher accuracy than the single-modality approaches.

Despite the progress made through these methods, problems remain for real-world solar yield prediction. One of the critical issues is that data of certain modalities can become missing in reality, leading to incomplete inference dataset. The aforementioned methods all assume that the data used for inference shares the same dimensions and completeness as the training data. However, in real-world scenarios, certain data input for deployed models could be easily lost due to network disconnection, insufficient bandwidth, hardware failure, battery constraints, or other unknown accidents. Such data missing is so prevalent that it can damage the prediction performance or even fail the model execution (as discussed in Section 5).

In this paper, we propose a novel **T**emporal **M**ulti-**M**odal **V**ariational **A**uto-**E**ncoder (TMMVAE) for solar power yield prediction, which highlights a more practical setting when input data is incomplete. For robustness to incomplete data, TMMVAE can impute the missing values in time-series sensor data and reconstruct them by consolidating multi-modality data. In specific, it incorporates 1) a mixture-of-experts variational auto-encoder module that synergizes different modalities of data for knowledge fusion and transfer, and 2) a temporal module that implements the temporal relationship within each modality to further enhance the imputation quality by considering the contexts. The results of extensive experiments on the real-world dataset show that our proposed network performs better than other multi-modality models across various levels of data incompleteness, including the scenario when one modality is completely missing. By leveraging the temporal multi-modal generative model, our system significantly improves the accuracy of missing data reconstruction. In consequence, it enhances the performance of downstream tasks for solar yield forecasting in both accuracy and robustness.

In summary, our main contributions could be listed as below:

- We propose a novel model with capability of imputing imperfect multi-modality time-series data. This model could perform better under various data missing settings.
- We build a dataset from our testbed, including solar energy yields, sky images, and meteorological data.
- We conduct extensive experiments on short-term forecasting of solar power generation to verify our proposed model and achieve more accurate and robust prediction results.

The remainder of the paper is organized as follows: Section 2 introduces the related works; Section 3 provides the overview and details of TMMVAE; Section 4 shows the data collection; Section 5 presents the evaluation results; and Section 6 summarizes the work and discusses the future research.

2 RELATED WORKS

In this section, we describe both physical and data-driven models applied in solar power forecasting, followed by an overview of the recent research works on data imputation techniques.

2.1 Solar Power Forecasting

There has been plenty of research works focusing on how to forecast short-term solar power generation. One group uses physical models [2, 13, 22] that employ the environmental monitoring data and the statistical information of solar panels, such as the power conversion efficiency and panel areas. However, these methods are found unreliable for short-term solar energy prediction [14, 20].

The other approach uses models driven by the numerical and visual data as shown in **Figure 1**. Numerical modality includes meteorological data collected from weather stations and solar power generation data acquired from solar panels. Statistical models are commonly used [16, 17, 21] on numerical modality to capture the relationship between historical and future solar generation. Visual modality consists of hemispherical sky images captured by ground-based fish-eye cameras to depict cloud floating direction, coverage, and other sky characteristics that can affect solar energy output. Deep learning-based techniques are often applied here to directly capture the relationship between visual information and trends of solar generation [23, 25, 32].

2.2 Data Imputation

The inference data can be missing from time to time due to various network or hardware accidents as illustrated in **Figure 1**. To strengthen the performance of solar power forecasting, it is necessary to impute the missing values first before we feed them into the prediction models. Existing solutions of data imputation could be roughly categorized as matrix factorization, variants of recurrent neural network (RNN), and generative models.

Temporal regularized matrix factorization [28] decomposes the multivariate time-series matrix into two smaller ones: the feature matrix and the latent matrix, for prediction tasks. The standard matrix completion technique is adopted to impute the missing values within the matrix.

Variants of RNNs are popular for imputing the missing values in time-series data [4, 5, 10, 27]. BRITS [4] directly uses a bi-directional RNN to predict incomplete data without particular assumptions or weight tuning. To fit with irregularly sampled data, the time decay factor proposed at [5] is incorporated into hidden state calculation. However, these methods are only applied on numerical time-series data imputation.

Another way to treat the data imputation problem is to make use of generative models. GAIN [26], a generative adversarial network-based method, uses a generator to recover the missing pieces of data and a discriminator to identify between imputed and true values. A hint mechanism is introduced, revealing partial information to the discriminator to reinforce the adversarial learning process. An alternative generative model, named variational auto-encoder (VAE), has also been applied to impute missing data. MIWAE [12] is a modification over importance-weighted auto-encoder (IWAE) to fit with missing-at-random data training and can thus impute incomplete data. HI-VAE [15] is a general framework of VAE to recover heterogeneous randomly missing observations by analyzing different continuous and discrete data distributions. To avoid Kullback-Leibler (KL) divergence loss dominating the whole evidence lower bound (ELBO) loss with incomplete training data, the ELBO loss is modified to solely depend on available observations.

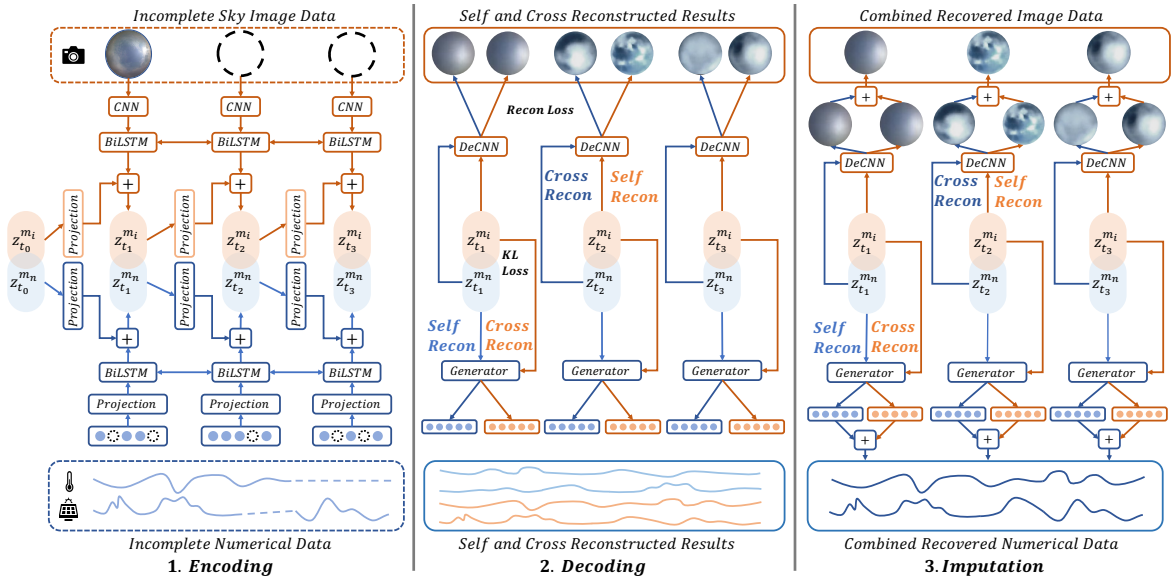


Figure 2: Architecture overview. The inference data could be randomly missing. We use dashed lines to represent the missing sky images or sensor data. To tackle this problem, we randomly delete training data and use a temporal mixture-of-experts of VAEs to encode image modality and numerical modality into their own latent spaces z^{m_i} and z^{m_n}, respectively in the first step. The numerical modality can then be self-reconstructed from z^{m_n} and cross-reconstructed form z^{m_i}. The same applies to the image modality. Next, we calculate KL loss and reconstruction loss to jointly optimize two experts. After training, we combine the self and cross reconstructed results into the final imputation depending on each modalities' completeness.

3 FRAMEWORK DESIGN

In this section, we first present the overview of the proposed architecture (as illustrated in **Figure 2**), and then describe each of the main parts in detail.

3.1 Problem Formulation

Mathematically, we define $X = \{x_{t,m}\}_{T,M}$ as the M modality time-series input within a time window size of T. In our case, there are two different modalities: 1) a numerical modality consisting of solar power generation data and meteorological information, as well as 2) a visual modality containing ground-based sky images.

We assume that our data is randomly missing without any specific patterns. We use X^O and X^U to represent observed and lost data, respectively, where $X^O \bigcap X^U = \emptyset$ and $X = X^O \bigcup X^U$. Our goal is to reconstruct the missing values $\hat{X}^U$ as close to the true values as possible. We then combine it with the observed raw data X^O to form a new complete input set $\hat{X} = \hat{X}^U \cup X^O$. This recovered input will be fed into the forecasting model for short-term solar energy yield prediction.

3.2 Architecture Overview

Our proposed framework aims to recover the missing inference data in order to strengthen the model for the downstream yield prediction task. Our network for data reconstruction utilizes information from all modalities and captures the temporal relationship within each modality for better data imputation quality. The imputation network consists of a mixture-of-experts variational auto-encoder with the temporal module that imputes the missing data within a time window. Using data retrieved from the imputation network, we employ a multi-modality model to finish the task of solar energy output prediction.

During training, we randomly delete half of observations in training data samples. The imputation network then generates the observations and the unseen parts through multiple VAEs. We calculate the KL loss of latent variables and the reconstruction loss of both self-reconstructed and cross-reconstructed data with ground truth. During inference, the model combines two imputation results over each modalities' completeness.

In the following subsections, we explain the design of each component and how we integrate them into the TMMVAE.

3.3 VAE for Mismatched Data Processing

In the first step, we modify the classical VAE and its training strategy to fit our scenario with missing data. In the classical VAE [8], we learn the distribution of input data $p(x)$ with the help of the latent variable z. To do so, we jointly train an inference model $q_\phi(z|x)$ to approximate the intractable posterior $p(z|x)$ and a generative model $p_\theta(z, x) = p_\theta(x|z)p_\theta(z)$ to map the sampled z to the corresponding x. To train a VAE model, we minimize the negative ELBO by applying gradient ascent with stochastic back-propagation on the parameters θ and ϕ:

$$\mathcal{L}(x) = KL[q_\phi(z|x)\|p_\theta(z)] - \mathbb{E}_{z\sim q_\phi(z|x)}[log p_\theta(x|z)]. \quad (1)$$

To tackle the missing data issue, i.e. $x = \{x^o, x^u\}$, we factorize our generative model and inference model accordingly. Our generative model $p_\theta(z, x) = p_\theta(x^o|z)p_\theta(x^u|z)p_\theta(z)$ gives both observed

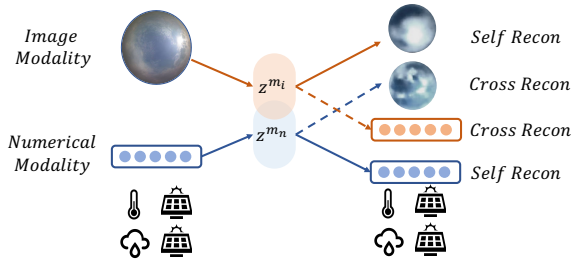


Figure 3: Mixture-of-Experts of VAEs. The z^{m_i} represents the latent space of image modality, and z^{m_n} represents the latent space of numerical modality. Dashed lines depict cross-reconstruction where data of one modality is generated from the latent space of the others. Solid lines on the right side represent the self-reconstruction of one modality from its own latent space.

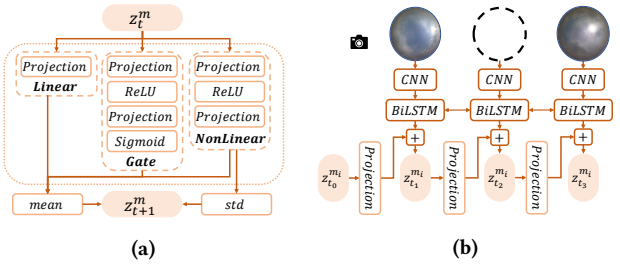


Figure 4: Temporal Module in VAE. (a) The structure of transition layer $p_\theta(z_t|z_{t-1})$ which computes the prior latent distribution recursively. (b) Our inference model $q_\phi(\vec{z}|\vec{x})$ which approximates the posterior latent distribution recursively.

and unobserved data. To ensure the generative model learns how to generate the unobserved information from observed data, we randomly delete parts of observations in the training stage and force the generative model to estimate the deleted values. For the inference model, the latent variable z is made to depend only on the observed attributes x^o since unobserved environmental variables will bring unavoidable noises. Following this intuition, we construct the inference model as $q_\phi(z|x) = q_\phi(z|x^o)$. We then rewrite the objective function as follows:

$$\mathcal{L}(x) = KL[q_\phi(z|x^o)\|p_\theta(z)] - \mathbb{E}_{z\sim q_\phi(z|x^o)}[log p_\theta(x^o|z)p_\theta(x^u|z)]. \tag{2}$$

The first term forces the inference model q_ϕ to make the posterior given observed inputs $q_\phi(z|x^o)$ close to the regular Gaussian prior $p_\theta(z)$. Meanwhile the second term encourages the generative model to reconstruct our observed and unobserved attributes which usually requires a richer posterior. This design confers the network the capability of missing value recovery.

3.4 Enhance VAE with Multi-Modal Learning

In the second step, we use the mixture-of-experts of VAEs to fuse multiple modalities. Mixture-of-experts variational auto-encoder [19] uses multiple VAE experts on multiple modalities where each expert learns the representation of its corresponding modality and maps it into the latent space. **Figure 3** illustrates the structure.

We first construct the inference model for modality m whose mission is to take x_m as its feeds and map them into the latent space, as in $q_\phi(z_m|x_m)$. We then modify the generative model into $p_{\theta_m}(z_m, x_{1:M}) = p_{\theta_m}(x_{1:M}|z_m)p_{\theta_m}(z_m)$, where $p_{\theta_m}(x_{1:M}|z_m)$ enables the generative model to reconstruct data from not only its own modality but also others'. This capability of both self- and cross-reconstruction is especially helpful when data of one modality is completely lost. We could use other experts to cross-reconstruct the entire missing modality. The objective function is obtained under missing data condition following equation 2:

$$\mathcal{L}(x) = \frac{1}{M}\sum_{m=1}^{M}\Big[KL[q_{\phi_m}(z_m|x_m^o)\|p_\theta(z_m)] - \mathbb{E}_{z_m\sim q_{\phi_m}(z_m|x_m^o)}[log p_\theta(x_{1:M}^o|z_m)p_\theta(x_{1:M}^u|z_m)]\Big]. \tag{3}$$

We average the loss among all the experts with the same assumption that each modality is comparable to others as in [19].

3.5 Leverage Temporal Information in VAE

In the third step, we leverage the temporal information in time-series data to improve data regeneration performance. Note that our observation x is a time sequence: $\vec{x} = \{x_1, x_2, \ldots, x_T\}$. To make use of temporal relations within observation sequences and simplify the temporal dependency structure, we take the concepts from state-space model [6, 9] and assume that our latent variable z is a sequence $\vec{z} = \{z_1, z_2, \ldots, z_T\}$ that follows Markov property and our observation x_t at a given time is conditioned only on the state z_t. Thus we could derive our generative model as:

$$p_\theta(\vec{x}, \vec{z}) = \prod_{t=1}^{T} p_\theta(x_t|z_t)p_\theta(z_t|z_{t-1}). \tag{4}$$

The decoders $p_\theta(x_t|z_t)$ of numerical modality and image modality are project layers and inverse convolution layers, respectively. To build the transition layer of latent variable $p_\theta(z_t|z_{t-1})$ in the generative model, we follow the design of gated transition layer in [9]. As shown in **Figure 4a**, the transition layer computes the distribution parameters of the next latent variable conditioned on the current latent variable and produces a gate component to control the weight of linearity. The mean of the next prior latent distribution is the weighted sum of linear and non-linear outputs and the standard deviation is non-linear output. By doing so, the generative model can capture both linear and non-linear transition relations in the latent space.

For inference model, we again follow the Markov property of latent sequence and decompose the approximated posterior. We then derive our inference model as:

$$q_\phi(\vec{z}|\vec{x}) = \prod_{t=1}^{T} q_\phi(z_t|z_{t-1}, \vec{x}). \tag{5}$$

To build such an inference model $q_\phi(\vec{z}|\vec{x})$, we follow the mature design of posterior approximation in [9]. We use a bi-directional LSTM network and feature extractor to transform the input sequence $\vec{x}$ into forward and backward hidden features $\vec{h}_f, \vec{h}_b$. Next, we use a projection layer followed by a Tanh activation layer to transform the last latent variable into feature $f_{z_{t-1}}$. We take the average of three features $\frac{1}{3}(h_{t,f}+h_{t,b}+f_{z_{t-1}})$ as the combined feature, and feed it into two projection layers which generates the mean and scale of posterior latent variables z_t. **Figure 4b** illustrates the module structure.

Following the aforementioned steps, we obtain our temporal variational auto-encoder that encodes the whole observation sequences first, and recursively computes the smoothing posterior $q_\phi(z_t|z_{t-1},\vec{x})$ subsequently, before reconstructing the observation sequences over the approximated posterior. Its objective function could be written as:

$$\mathcal{L} = \sum_{t=1}^{T} \Big[KL[q_\phi(z_t|z_{t-1},\vec{x}) \| p_\theta(z_t|z_{t-1})] - \mathbb{E}_{z\sim q_\phi(z_t|z_{t-1},\vec{x})}[log p_\theta(x_t|z_t)] \Big]. \tag{6}$$

3.6 Temporal Multi-Modal Variational Auto-Encoder (TMMVAE)

In the last step, we develop our final model that fits with incomplete data and utilizes multi-modality and temporal information.

In this model, we have two modalities: image modality and numerical modality, referred to as m_i and m_n, respectively. At time index of t, the inference model $q_{\phi_m}(\vec{z}|\vec{x})$ absorbs the encoded feature of the whole observed sequences $\vec{x}^o_m$ and the latent variable from the last step $z_{m,t-1}$ to generate the posterior of latent variables. Then the generative model p_{θ_m} computes the prior of $z_{m,t}$ from the last step $z_{m,t-1}$ for KL divergence loss computation. Now we have two latent spaces, z_{m_i} for image modality and z_{m_n} for numerical modality. We sample from each of the two spaces and pass results to $p_{\theta_{m_i}}(x_{m_i,t}|z_{m_i,t})$ for reconstructing the observed and unobserved images. We then pass the same samples to $p_{\theta_{m_n}}(x_{m_n,t}|z_{m_n,t})$ for reconstructing the observed and unobserved numerical data. Note that if the modality of reconstructed data is the same as that of the latent space derived from the corresponding inference model, we call this regeneration process self-reconstruction; Otherwise, we call it cross-reconstruction. The objective function can be written according to equations (2,3,6):

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T} \Big[KL[q_{\phi_m}(z_{m,t}|z_{m,t-1},\vec{x}^o_m) \| p_{\theta_m}(z_{m,t}|z_{m,t-1})] - \mathbb{E}_{z_{m,t}\sim q_{\phi_m}(z_{m,t}|z_{m,t-1},\vec{x}^o_m)}[log p_\theta(x^o_{1:M,t}|z_{m,t}) p_\theta(x^u_{1:M,t}|z_{m,t})] \Big]. \tag{7}$$

During the training process, we randomly mask out parts of training data to mimic data missing but with ground truth so that we could calculate the reconstruction error for unobserved data. This masking technique also helps generalize the model to work well under different degrees of data incompleteness by capturing the pattern of both observed and unobserved information. The whole training process is described by **Algorithm 1**.

Algorithm 1: Training Process of TMMVAE

Input: Training dataset X including numerical data x_{m_n} and visual data x_{m_i}, hyper-parameters: learning rate, weight decay, time window size T;
Output: Model parameters ϕ, θ ;
while *epoch* < *MaxEpochs* **do**
  Sample $\vec{x}_{m_i}, \vec{x}_{m_n}$ of time length T from dataset ;
  Randomly mask half of data points and images ;
  Generate z_{m_n} from $q_{\phi_{m_n}}$;
  for m *in* $[m_i, m_n]$ **do**
    Sample $\vec{z}_m$ from $q_{\phi_m}(z_t|z_{t-1},\vec{x}^o_m)$ recursively;
    Sample $\vec{z}'_m$ from $p_{\theta_m}(z_t|z_{t-1})$ recursively;
    Compute KL loss of $\vec{z}_m$ and $\vec{z}'_m$;
    Reconstruct $\vec{\hat{x}}_m$ from p_{θ_m} and $\vec{z}_m$;
    Compute NLL self reconstruction loss ;
    if $m == m_i$ **then**
      Cross reconstruct $\vec{\hat{x}}_m$ from p_{θ_m} and $\vec{z}_{m_n}$;
    else
      Cross reconstruct $\vec{\hat{x}}_m$ from p_{θ_m} and $\vec{z}_{m_i}$;
    end
    Compute NLL cross reconstruction loss ;
  end
  Accumulate loss and update parameters ϕ, θ ;
  epoch += 1;
end

3.7 Reconstruct the Missing Modality

To recover the lost data in the inference stage, we adopt the proposed TMMVAE. Take numerical data as an example. First, we embed our captured image data and numerical sensor data into their own latent spaces. We then derive two samples z_i and z_n from the latent spaces of image and numerical modalities, respectively. We can both self-reconstruct the missing numerical data from samples z_n of the same modality and cross-reconstruct them from samples z_i of the image modality. These two reconstructed outputs are combined as final imputation results.

A few points are worth noting. Firstly, when the missing ratio and pattern of two modalities are the same, the accuracy of self-reconstruction is usually higher than that of cross-reconstruction. Secondly, when one modality of data is corrupted seriously while other modalities are fine, we could impute the severely corrupted modality from others, which provides better imputation quality than the self-reconstruction process. Following these findings, we combine the reconstruction results over the completeness of each modality, c_m, as:

$$\hat{x}_{m,t} = \begin{cases} x_{m',t} & , c_{m,t} = 0 \\ x_{m,t} & , c_{m,t} > 0 \end{cases} (m' \neq m). \tag{8}$$

3.8 Forecasting Tasks

We build several deep-learning-based models for short-term solar energy prediction. To demonstrate the effectiveness of multi-modal learning, we first build single-modality algorithms using either

visual or numerical modality. For single visual modality methods, we build a CNN-LSTM-based network as the predictive model. And for numerical modality methods, we employ two networks that are either LSTM-based or TCN-based [1]. We then build a multi-modality model similar to the model used at [31] and we name it as MM-Pre. MM-Pre incorporates a CNN-LSTM module extracting the context of image sequence, a LSTM module extracting the context of numerical sequence, and several projection layers transforming the concatenated context into predictions.

4 DATA COLLECTION

We collect data from weather station sensors, solar panels and ground-based sky cameras as described in **Figure 1**. Specifically, three types of data are collected for solar power yield forecasting:

Solar power generation data. We obtain power generation data by reading the measurements from smart meters attached to the solar plants. Data from 32 active solar plants are used in total. The update frequency is 30 minutes.

Meteorological data. To collect rich and accurate weather information, we set up a Davis weather station (Vantage Pro2) on site. The obtained real-time meteorological data has 6 attributes including temperature, humidity, dew point, rain rate, UV index, and solar radiation. The update frequency is 1 minute.

Sky images. Ground-based wide-angle cameras could capture hemispheric sky images that cover a large area. Compared with the weather data, sky images have two advantages: 1) they could enable visual analysis on cloud and sun by providing information such as cloud coverage and sun position sun; 2) they could cover a wider physical range than weather data which only records single-point environment. These two strengths render sky images particularly useful in short-term solar yield prediction. We obtain the sky images from our collaborator and the update frequency is 1 minute.

To align these information, we average the meteorological data within a time window of 30 minutes, and choose the sky image taken in the middle of the time window. The data missing percentage of each attribute in dataset is shown in **Figure 5**. We collect around five months of data with 5156 data samples in total.

5 EXPERIMENTS

In this section, we first introduce our evaluation metrics and the compared methods for data imputation. We then describe in detail the experiments and discuss their results. Finally, we evaluate the performance of various models applied in forecasting tasks.

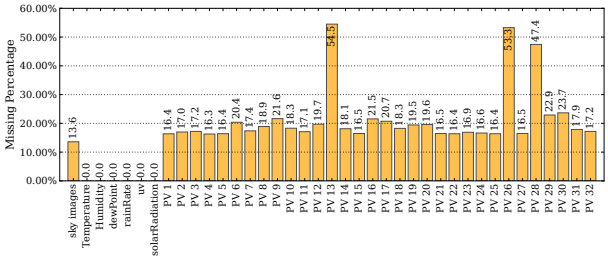


Figure 5: Data missing percentage of each attribute in our self-collected dataset.

5.1 Evaluation Metrics

We compare methods for data recovery with respect to the error of data imputation. In numerical modality, there are multiple domains of data collected, including UV index, rain rate, and power generation from solar panels. The magnitude and unit of these domains can vary, so we compute the average relative imputation error over all domains within each modality. We choose AvgNRMSE for error evaluation, calculated as:

$$AvgNRMSE = \frac{1}{F}\sum_{f}\left(\frac{\sqrt{\frac{1}{n}\sum_{n}(x_{n,m_f} - \hat{x}_{n,m_f})^2}}{max(x_{m_f}) - min(x_{m_f})}\right), \tag{9}$$

where F is the total number of features within modality m , x is the ground truth, and $\hat{x}$ is the imputed value.

5.2 Compared Imputation Methods

We include the methods used for data imputation as below:

Mean. This is to simply fill in the missing data using the average value of the remaining data within the same modality.

GAIN. This is a generative adversarial model including a discriminator to identify the imputed data from the rest, and a generator to estimate the true value of the missing data [26].

TMVAE. MVAE is a variational auto-encoder that learns the joint posterior distribution through the product-of-experts [24]. Following the assumption that the marginal posterior distribution is Gaussian, the product of Gaussian will be more influenced by data with higher confidence, or in other words, with lower variance. We add the same temporal module mentioned in section 3.5 into MVAE to fit it for time-series data.

TVAE-Num. This is a temporal VAE network that only takes numerical modality data as its input. We use this network for comparison to show that imputation accuracy could be improved by learning from both image and numerical modalities.

MMVAE. This is a multi-modal VAE with mixture-of-experts but without temporal modelling. We compare this model with TMMVAE to demonstrate the effectiveness of embedded temporal module.

TMMVAE. Our proposed model builds two experts of VAE together with temporal module to handle numerical modality and image modality respectively and optimizes them jointly.

5.3 Implementation Details

Our model uses time window of two and a half hours, i.e., the input sequence length is 5 units long. For feature extractor of visual modality, we use 5 layers of CNN where each layer is followed by a Batch Normalization layer and a LeakyReLU activation layer. For feature extractor of numerical modality, we use one linear layer without bias to first encode the numerical data at each time step followed by a LeakyReLU activation layer and a non-linear layer with bias. The first linear layer without bias helps avoid taking zero-imputed values as inputs. In each expert, we use a Bi-LSTM network to encode these features. The transition layer of latent space contains several projection layers. The Bi-LSTM output and transition output are combined as a smoothing feature. We use two projection layers to project this feature to mean and std of the latent variable, respectively. We then sample z from mean and std variables by applying the reparameterization technique.

Models	Input Modality		Numerical Modality Missing Percentage									
	Visual	Numerical	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Mean	×	✓	0.0405	0.0678	0.0976	0.1239	0.1546	0.1870	0.2185	0.2533	0.2889	0.3252
GAIN	×	✓	0.0597	0.0580	0.0575	0.0586	0.0615	0.0661	0.0725	0.0816	0.1270	0.2969
TVAE-Num	×	✓	0.0430	0.0380	0.0372	0.0469	0.0707	0.1028	0.1372	0.1744	0.2046	0.1909
TMVAE	✓	✓	0.0428	0.0402	0.0391	0.0385	0.0387	0.0423	0.0536	0.0807	0.1316	0.2522
MMVAE-Self	✓	✓	0.0360	0.0355	0.0370	0.0387	0.0421	0.0531	0.0756	0.1151	0.1724	0.2472
MMVAE-Cross	✓	✓	0.1386	0.1386	0.1386	0.1386	0.1386	0.1386	0.1386	0.1386	0.1386	0.1386
MMVAE-Full	✓	✓	0.0360	0.0355	0.0370	0.0386	0.0420	0.0530	0.0755	0.1148	0.1688	0.1386
TMMVAE-Self	✓	✓	**0.0312**	**0.0317**	**0.0323**	**0.0329**	**0.0340**	**0.0352**	**0.0371**	**0.0410**	**0.0523**	0.2604
TMMVAE-Cross	✓	✓	0.1233	0.1233	0.1233	0.1233	0.1233	0.1233	0.1233	0.1233	0.1233	**0.1233**
TMMVAE-Full	✓	✓	**0.0312**	**0.0317**	**0.0323**	0.0330	**0.0340**	**0.0352**	**0.0371**	0.0421	0.0533	**0.1233**

Table 1: The AvgNRMSE of numerical data imputation using different methods under different data missing percentage. We randomly remove numerical data points while preserving image data. The best results for each column are highlight in bold.

We normalize the numerical data to zero mean and unit variance. The training batch size is 64. We use an Adam optimizer with a learning rate of 0.001 and a step-wise learning rate scheduler to divide the learning rate by 10 at epoch 50 and 100. The maximum epoch of training is set as 200, and we save the top-10 models on numerical data evaluated by the metrics of imputation error.

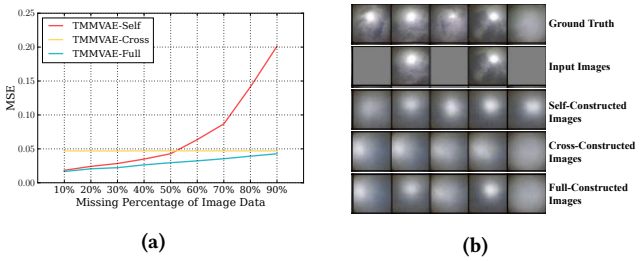
5.4 Overall Results

To comprehensively demonstrate the model performance on missing data recovery, we compare the performance of all approaches across different data missing percentages. In the first experiment, we only randomly remove parts of the numerical modality and keep the image modality complete. Then we show the AvgNRMSE value on all domains of numerical modality for all methods in **Table1**.

First, we compare the methods that only take numerical modality as input. GAIN, one of the GAN-based methods, fails to impute missing data well on corrupted dataset. Although there are more advanced methods, such as GRUI [11], that could potentially outperform GAIN on time-series data, we argue that GAN-based methods are more unstable and harder to train compared to VAE-based methods, especially under the unstructured multi-modality data settings. TVAE-Num could outperform GAIN when numerical data misses less than 50%, which reflects the potential capability of VAE on imputation tasks.

Next, we compare the methods that take data from both numerical and visual modalities. TMVAE, employing the product-of-experts to fuse multiple modalities, fails to outperform TVAE-Num with single modality. We attribute this performance degradation to the mechanism of product-of-experts which leads to a latent space dominated by visual modality when numerical modality is missing. MMVAE is the mixture-of-experts VAE without the temporal module, and thus with massively missing data, the model loses its ability to impute data from available information.

TMMVAE-self represents the model that reconstructs missing numerical data from observed data of the same modality, while TMMVAE-cross represents the model that reconstructs missing numerical data from observed sky images. Interestingly, even if we preserve sky images and delete only the numerical data, the cross-reconstruction of numerical data from images still performs worse than the self-reconstruction approach. This is because only limited information can be extracted from sky images to represent the corresponding numerical data. However, when numerical modality is completely missing, visual modality appears helpful to cross-reconstruct its data. As results, TMMVAE-Full which combines both self- and cross-reconstructed data could impute more robustly compared to TMMVAE-Self and TMMVAE-Cross models.



(a) (b)

Figure 6: Sky images reconstruction. (a) The MSE of image reconstruction under different image missing percentages. (b) Reconstruction of half missing sky images.

In the second experiment, we evaluate model performance on image reconstruction by randomly deleting images while preserving numerical data. By combining the self-reconstructed images and those cross-reconstructed from numerical modality, TMMVAE-Full achieves lower reconstruction mean-square-error (MSE) as shown in **Figure 6a**. **Figure 6b** illustrates the reconstruction results of half missing sky images (deleted images are shown in grey in the input sequence).

5.5 Sensitivity Analysis

To further analyse the model performance in other cases, we conduct sensitivity analysis. In the first experiment, we delete data from the numerical modality at the block level randomly while preserving those from the visual modality. Different from the last experiment where the deletion of data points is completely random, this experiment will delete all data points collected at time t as a block. We show the results of imputation error on numerical data

Models	Numerical Modality Missing Percentage				
	10%	30%	50%	70%	90%
GAIN	0.1143	0.1813	0.2122	0.2490	0.2833
TVAE-Num	0.0904	0.1344	0.1714	0.1996	0.2341
MMVAE-Self	0.0912	0.1404	0.1789	0.2066	0.2342
MMVAE-Full	0.0600	0.0858	0.1049	0.1198	0.1329
TMMVAE-Self	0.0828	0.1299	0.1688	0.1998	0.2391
TMMVAE-Full	**0.0531**	**0.0771**	**0.0954**	**0.1088**	**0.1190**

Table 2: The AvgNRMSE of numerical data imputation with various data missing percentages. We remove the numerical data block by block randomly while preserving the image data. The best result in each column is highlighted in bold.

Models	Modality Missing Percentage				
	10%	30%	50%	70%	90%
TMVAE	0.0662	0.1180	0.1626	0.2035	0.2395
MMVAE-Self	0.0912	0.1404	0.1789	0.2066	0.2342
MMVAE-Full	0.0647	0.1043	0.1455	0.1853	0.2276
TMMVAE-Self	0.0828	0.1299	0.1688	0.1998	0.2391
TMMVAE-Full	**0.0538**	**0.0800**	**0.1082**	**0.1447**	**0.2139**

Table 3: The AvgNRMSE of numerical data imputation with different missing percentages. We remove both numerical data and image data block by block randomly. The best result in each column is highlighted in bold.

in **Table 2**. Compared to the results in Table 1, the imputation error under the same missing percentage increases since blocked missing data is harder to impute. When more than half of the data blocks are missing, GAIN and TMMVAE-self methods fail to perform well, suffering from insufficient information, while TMMVAE-full achieves better imputation quality by leveraging data from image modality.

We conduct another experiment with data deletion on both visual and numerical modalities on the block level with the same percentage steps, and report the imputation error of numerical data in **Table 3**. As shown, TMMVAE-full performs more strongly than other multi-modal VAE-based methods when data of both modalities are missing.

5.6 Forecasting Tasks

Herein, we first evaluate single- and multi-modality models mentioned at Section 3.8 on the task of predicting solar power generation in next half hour. CNN-LSTM model intakes image modality data only, LSTM- and TCN-based models use numerical modality data only, and MM-Pre model utilizes data from both modalities. As shown in **Table 4**, the multi-modal learning-based method MM-Pre achieves the best prediction accuracy compared to single-modality methods, indicating that sky images are beneficial for short-term solar energy forecasting.

We then evaluate the data imputation performance of different models by comparing the forecasting error of short-term yields when data is missing to various degrees. We randomly remove both numerical data and image data. **Figure 7** illustrates that the

Models	Input Modality		Prediction Error
	Visual	Numerical	
CNN-LSTM	✓	×	0.1986
LSTM-based	×	✓	0.0933
TCN-based	×	✓	0.0861
MM-Pre	✓	✓	**0.0826**

Table 4: Four models evaluated for short-term solar yield prediction. MM-Pre achieves the best prediction accuracy.

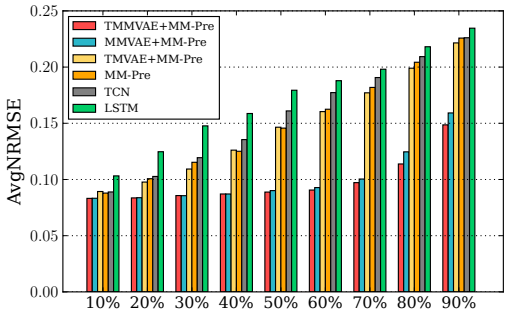


Figure 7: Forecasting AvgNRMSE by different models with various data missing percentages. Both numerical data and image data are randomly removed.

forecasting accuracy of the three methods (we do not consider CNN-LSTM model here since it performs much worse than others) without any data imputation module drops dramatically. The other three methods reconstruct the missing parts first before forecasting. Benefiting from the imputed data, these forecasting models perform more robustly with severely corrupted inference data. Among these results, TMMVAE delivers the most stable and accurate performance.

6 CONCLUSION

In this paper, we propose a new method for data imputation on multi-modality time-series data, called temporal multi-modal variational auto-encoder. In the application of short-term solar energy yield forecasting, the sensor data may become missing from time to time, leading to prediction performance degradation. With the proposed TMMVAE network, we could recover both missing numerical sensor data and lost sky images, hence providing more accurate and robust short-term solar yield prediction. This will in turn enhance the stability of the modern power grid where renewable energy plays an increasingly significant role.

ACKNOWLEDGMENTS

This work is supported by the National Research Foundation, Singapore, the Energy Market Authority, under its Energy Programme (EP Award <NRF2017EWT-EP003-023>), and MOE under its grant call (RG96/20).

REFERENCES

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[2] Kilian Bakker, Kirien Whan, Wouter Knap, and Maurice Schmeits. 2019. Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation. *Solar Energy* 191 (2019), 138–150.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, Vol. 31.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 1–12.

[6] Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Theophane Weber. 2019. Temporal Difference Variational Auto-Encoder. In *International Conference on Learning Representations*.

[7] IRENA 2021. International Renewable Energy Agency's analytics portal of solar energy. https://www.irena.org/solar. Accessed: 2021-04-04.

[8] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[9] Rahul Krishnan, Uri Shalit, and David Sontag. 2017. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[10] Zachary C Lipton, David C Kale, Randall Wetzel, et al. 2016. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare* 56 (2016).

[11] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, and Xiaojie Yuan. 2018. Multivariate time series imputation with generative adversarial networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1603–1614.

[12] Pierre-Alexandre Mattei and Jes Frellsen. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*. PMLR, 4413–4423.

[13] Sthitapragyan Mohanty, Prashanta K. Patra, Sudhansu S. Sahoo, and Asit Mohanty. 2017. Forecasting of solar energy with application for a growing economy like India: Survey and implication. *Renewable and Sustainable Energy Reviews* 78 (2017), 539–553.

[14] Akinobu Murata, Hideaki Ohtake, and Takashi Oozeki. 2018. Modeling of uncertainty of solar irradiance forecasts on numerical weather predictions with the estimation of multiple confidence intervals. *Renewable Energy* 117 (2018), 193–201.

[15] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition* 107 (2020), 107501.

[16] Hossein Panamtash, Qun Zhou, Tao Hong, Zhihua Qu, and Kristopher O Davis. 2020. A copula-based Bayesian method for probabilistic solar power forecasting. *Solar Energy* 196 (2020), 336–345.

[17] Mashud Rana, Irena Koprinska, and Vassilios G Agelidis. 2016. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management* 121 (2016), 380–390.

[18] Walter Richardson, Hariharan Krishnaswami, Rolando Vega, and Michael Cervantes. 2017. A low cost, edge computing, all-sky imager for cloud tracking and intra-hour irradiance forecasting. *Sustainability* 9, 4 (2017), 482.

[19] Yuge Shi, Siddharth N, Brooks Paige, and Philip Torr. 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

[20] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd. Rahim. 2018. Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management* 156 (2018), 459–497.

[21] William VanDeventer, Elmira Jamei, Gokul Sidarth Thirunavukkarasu, Mehdi Seyedmahmoudian, Tey Kok Soon, Ben Horan, Saad Mekhilef, and Alex Stojcevski. 2019. Short-term PV power forecasting using hybrid GASVM technique. *Renewable energy* 140 (2019), 367–379.

[22] Cyril Voyant, Marc Muselli, Christophe Paoli, and Marie-Laure Nivet. 2012. Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. *Energy* 39, 1 (2012), 341–355.

[23] Fei Wang, Zhiming Xuan, Zhao Zhen, Yu Li, Kangping Li, Liqiang Zhao, Miadreza Shafie-khah, and João PS Catalão. 2020. A minutely solar irradiance forecasting method based on real-time sky image-irradiance mapping model. *Energy Conversion and Management* 220 (2020), 113075.

[24] Mike Wu and Noah Goodman. 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *NeurIPS*. 5580–5590.

[25] Handa Yang, Ben Kurtz, Dung Nguyen, Bryan Urquhart, Chi Wai Chow, Mohamed Ghonima, and Jan Kleissl. 2014. Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego. *Solar Energy* 103 (2014), 502–524.

[26] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*. PMLR, 5689–5698.

[27] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. 2017. Multi-directional recurrent neural networks: A novel method for estimating missing data. In *Time Series Workshop in International Conference on Machine Learning*.

[28] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. 2016. Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.

[29] Huaizheng Zhang, Han Hu, Guanyu Gao, Yonggang Wen, and Kyle Guan. 2018. DeepQoE: A unified framework for learning to predict video QoE. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[30] Huaizheng Zhang, Yong Luo, Qiming Ai, Yonggang Wen, and Han Hu. 2020. Look, read and feel: Benchmarking ads understanding with multimodal multitask learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 430–438.

[31] Jinsong Zhang, Rodrigo Verschae, Shohei Nobuhara, and Jean-François Lalonde. 2018. Deep photovoltaic nowcasting. *Solar Energy* 176 (2018), 267–276.

[32] Zhao Zhen, Jiaming Liu, Zhanyao Zhang, Fei Wang, Hua Chai, Yili Yu, Xiaoxing Lu, Tieqiang Wang, and Yuzhang Lin. 2020. Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image. *IEEE Transactions on Industry Applications* 56, 4 (2020), 3385–3396.