Multi-Branch ResNet-Transformer for Short-term Spatio-Temporal Solar Irradiance Forecasting

Saeedeh Ziyabari, Zhenyu Zhao, Liang Du, *Senior Member, IEEE*, Saroj K. Biswas *Life Senior Member, IEEE*

***Abstract*—The increasing penetration of solar generation into power grids has promoted the need for accurate and reliable short-term solar irradiance forecasting. Existing methods utilizing advanced deep learning architectures have shown advanced performance compared to conventional time-series analytical techniques but in general encountered shortcomings in modeling spatial correlations among neighboring solar generation sites, exploring the similarity of long-term, time-varying patterns, and alleviating overfitting issues in convolutional and recurrent neural networks, such as the popular Long Short-term Memory (LSTM). To effectively but yet reliably tackle these challenges in the existing literature, this paper proposes a spatio-temporal framework consisting of a multi-branch hybrid Residual network and the `Transformer` architecture (ResTrans). The proposed framework has been tested on two groups' real-world data containing 17 years-long data from different solar sites in Philadelphia, USA, including 12 and 18 locations, respectively. Compared to other hybrid benchmark architectures, including single-branch ResTrans and multi-branch ResNet-LSTM (ResLSTM), single-branch ResLSTM, and CNN-LSTM, the proposed multi-branch ResTrans achieves the highest forecasting accuracy with an average RMSE of 0.049 (W/m^2), an average MAE of 0.031 (W/m^2), and an R^2 coefficient of 97%.**

***Index Terms*—solar irradiance forecasting, spatio-temporal modeling, deep residual network (ResNet), attention mechanism, transformer neural network**

NOMENCLATURE

Indices

n	index of solar generation sites
k	index of input feature variables
i	index of residual blocks
j	index of branches
m	index of layers within residual block

Parameters

I	total number of residual blocks
J	total number of branches
K	total number of input feature variables
N	total number of solar generation sites
γ	learning rate of multi-branch ResNet
a_t	attention score at time t
h'_t	attention vector of hidden layer
h_n	n^{th} head of multiple-head self-attention
L	window length
$C(\tau)$	time embedding
Q, K, V	query, key, and value matrices of self-attention
h_t, p_t	hidden layer and output vector
$x_t^{n,k}$	k-th input variable at solar site s^n at time t
X_t^n	all input variables at solar site s^n at time t
W_h	weight of hidden layer
$Y_t^n, \hat{Y}_t^n$	measured / forecast GHI values at s^n at time t
g, ω, φ	periodic feature and two learnable parameters
Δ	forecasting horizon
Z^o	learned projection matrix
Z_0, Z_D	input and output of the residual network
$Z_{i,j}$	input of i-th residual block and j-th branch
$\Phi_{i,j}$	weight w.r.t. i-th residual block and j-th branch
$\bar{z}$	mean of the variable
R^2	statistical measurement of prediction model

Functions

$\sigma(\cdot)$	sigmoid function
$\mathcal{L}(\cdot)$	loss function
$G(\cdot)$	residual function
$\psi(\cdot)$	softmax function
$F(\cdot)$	forecast function

Manuscript received January 30, 2023; revised April 30, 2023 and May 10, 2023; accepted June 7, 2023. Date of publication XXX, 2023; date of current version XXX, 2023. L. Du was supported in part by the National Science Foundation (NSF) under Award 2238414 and by the National Academies of Science, Engineering, and Medicine (NASEM) under Gulf Research Program's Early-Career Research Fellowship. Paper no. 2023-ESC-0038. (Corresponding author: Liang Du.)

The authors are with the Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122, USA. E-mail:{saeedeh, z.zhao, ldu, sbiswas}@temple.edu.

This paper was presented at 2022 IEEE Energy Conversion Congress and Exposition (ECCE) [1].

I. INTRODUCTION

In recent years, the widely proliferation of distributed energy resources (DERs) has caused major paradigm shifts for future electric power grids. For instance, the recent U.S. Federal Energy Regulatory Commission (FERC) Order 2222 allows DERs to participate in regionally organized wholesale electricity markets and compete with conventional thermal generators [2]. Consequently, system operators and aggregators face challenges in how to effectively handle uncertainty and volatility in high penetration of DERs, especially ubiquitous solar generations that could be before- or behind-the-meter [3], [4]. Therefore, daily operations of both electricity markets and vertical integrations requires granular and reliable awareness of solar generation capability, which in term relies on accurate and robust solar irradiance forecasting mechanisms at different time scales [5]. For instance, studies have shown significant economic benefits utilizing intelligent self-scheduling based on reliable sky images-based solar irradiance forecasting techniques [6].

In general, solar irradiance is considered a function of highly volatile meteorological factors at various time

scales [7]. In a prediction horizon manner, solar irradiation forecasting techniques can be classified into four categories based on their target forecast time ahead: (i) very short, (ii) short, (iii) medium, and (iv) long-terms, and a detailed comparison can be found in [5]. Specifically, short-term forecasting, typically ranging from several hours to one day ahead, is highly challenging due to volatile short-term fluctuations in solar irradiance, mostly driven by local meteorological conditions [8], [9]. Therefore, this paper will be focused on short-term solar irradiance forecasting.

Moreover, in practical applications, solar irradiance is forecasted in either a temporal or spatio-temporal manner. Specifically, temporal techniques predict the solar irradiance based on historical data on a specific site, whereas spatio-temporal methods utilize historical data from multiple sources with inherent spatial correlations. If handled properly, spatial correlations among data sources (i.e., the spatio-temporal approach) could lead to significant improvements in prediction accuracy compared to the temporal method [10]. Compared to the conventional spatio-temporal solar irradiance forecast literature that is based on data-driven time-series data analysis, machine learning (ML)-based models have show better performance. Specifically, Deep Learning (DL) models includes Recurrent Neural Networks (RNNs) and convolutional neural networks (CNNs) have been widely adopted for temporal and spatial modeling, respectively.

Furthermore, for better performances, more sophisticated ML frameworks are adopted not only for forecasting at spatio-temporal scales and also incorporating inherent non-linear characteristics of solar irradiance data [11], [12]. For instance, in [13], the authors developed a hybrid architecture that adopt long short-term memory (LSTM) to extract temporal characteristics and use CNN to learn spatial features, respectively. In [11], a novel neural network autoencoder is proposed as a generative probabilistic model to learn the continuous probability densities of each solar site's data. The authors adopt graph spectral convolutions to capture the spatial characteristics of each site, the features then forwarded to an encoder and decoder neural network to learn the distribution of solar irradiance and forecast. In [14], a framework is proposed to combine graph CNN and LSTM to process the signal with a graph perspective in a spatio-temporal manner to achieve better performance on forecast solar irradiance.

Compared to CNN, aforementioned RNN-based frameworks are capable of extracting temporal dependencies, however, have certain limitations when dealing with long sequence data. When input sequences become longer, RNNs become computationally inefficient and cause gradient vanishing. Also, the RNNs lack the ability to learn long-term dependencies in a long sequence [15]. Although the LSTM-based frameworks can alleviate the gradient vanishing problem and learn both short and long-term dependencies, LSTM-based models have limitations in remembering long sequences [16].

To overcome the shortcomings mentioned, this paper proposes a novel multi-branch ResTrans architecture that combines a `Transformer` network for temporal modeling and a multi-branch residual network (ResNet) for spatial modeling. The `Transformer` neural network is first proposed in [17] for overcoming the long-range dependencies problem in sequence-to-sequence tasks in Natural Language Processing (NLP). The `Transformer`'s architecture is solely based on attention mechanisms. That means `Transformer` is free from recurrence and convolutions entirely and overcomes all shortcomings with them [17]. Beyond the application in NLP, the framework based on the `Transformer` with attention mechanism has also been adopted for research involving time-series data. In [18], the authors propose a `Transformer`-based GAN network to solve the mode collapse problem in time-series anomaly detection (TSAD) and improve the generalization capability. In [19], a convolutional neural network with `Transformer`-encoder is proposed. Compared to common approaches with convolution neural network with long and short-term memory structures, the framework with `Transformer`-encoder mine the deep information in multivariate time series more accurately. In [20], a multi-head `Transformer` framework that is `Transformer`-based is proposed to forecasts the patient's status time series variables utilizing various vitals signs of the patient to capture the long-term dependencies.

Inside the proposed framework, the `Transformer` network based on attention mechanisms, is developed to enable parallel processing, minimize information loss, and alleviate the inefficiencies caused by recursion and sequential processing [17]. The primary benefits the proposed framework brings are listed as follow:

- The multi-branch ResNet leverages feature extraction at multiple time scales by applying several convolutional branches
- The over-fitting problem is alleviated by exploiting shared representations as auxiliary information
- The learning process is accelerated
- The `Transformer` analyzes the long sequences data more quickly and efficiently

To the best of our knowledge, the proposed framework in this paper is the first work that combines `Transformer` with multi-branch ResNet networks for short-term spatio-temporal solar irradiance forecasting. The performance of our proposed framework is tested on real data from different solar sites in Philadelphia, USA. The proposed framework outperforms mainstream benchmark architectures, including single-branch ResTrans and multi-branch ResNet-LSTM (ResLSTM), single-branch ResLSTM, and CNN-LSTM on two different data groups containing 12 sites (validation of temporal models) and 18 sites (validation of spatio-temporal models), respectively [21].

The remaining of this paper is organized as follows. Section II formulates the proposed short-term solar forecast problem and outlines the proposed framework with necessary background information. Sections III proposes temporal solar forecast models based on the `Transformer` architecture with time embedding steps to process historical solar irradiance data. Furthermore, Section IV further extends to spatio-temporal solar forecast models based on `Transformer` and ResNet. Section V describes the data and its process for numerical validations, which are presented in Section VI and

compared with the literature to validate the proposed models. Finally, Section VII draws conclusions and summarizes contributions by this paper.

II. Problem Formulation and Background

This section outlines the proposed framework, which consists of `Tansformer`, attention mechanism, and residual network (ResNet).

A. Problem Definition

We follow our previous work [5] and formally define the considered short-term solar irradiance forecast problem as follows. Consider a set of solar generating sites$S = \{s^1, s^2, \ldots, s^N\}$, where s^n denotes the n^{th} solar site and is associated with an input vector $X_t^n = [x_t^{1,n} \ldots x_t^{k,n} \ldots x_t^{K,n}]$ of K meteorological variables and its corresponding actual values $Y_t^n \in \mathbb{R}$ at each time step $t \in [1, T]$. Consequently, the predicted solar irradiance value at s_n over a defined forecasting horizon Δ is estimated by

$$\hat{Y}_{t+\Delta}^n = F(X) \tag{1}$$

where $F(\cdot)$ donates the forecasting function and $\hat{Y}_{t+\Delta}^n$ is the estimated value of the actual $Y_{t+\Delta}^n$. In this work, $F(\cdot)$ is implemented by a hybrid architecture consisting of the self-attention mechanism, ResNet, and the `Transformer`. The fundamental architectures of these functional models are discussed as follows.

B. The `Transformer`

As a primary part of the proposed framework, the `Transformer` is a transduction model proposed in [17] which utilizes the self-attention mechanism to compute representations of its inputs and outputs without using sequence-aligned RNNs or convolution. The `Transformer` is widely adopted to replace models that are based on the conventional encoder-decoder architecture, such as LSTM and GRU, which have encountered various difficulties in handling sequence-to-sequence tasks with long-range dependencies, as discussed in Section I.

As shown in Fig. 1, the original `Transformer` model consists of multiple self-attention modules and time-embedding to identify periodic elements in the variation of time-series data and non-periodic elements for output prediction. Each encoder consists of a multi-head self-attention mechanism and a fully connected position-wise feed-forward network. The decoder is similar to an encoder with one extra multi-head self-attention over the encoder's outputs. Moreover, the `Transformer` can be constructed as a stack of multiple encoder and decoder building blocks, each of which works independently without the need of sharing weights. The model's outputs are generated based on the encoded input data and previous decoder outputs [17]. As the RNNs are designed to handle sequential inputs such as time series data [15], such a stacked `Transformer` model generally outperform RNNs for sequential inputs. Moreover, with the attention mechanism added, the `Transformer` can compute inputs and outputs

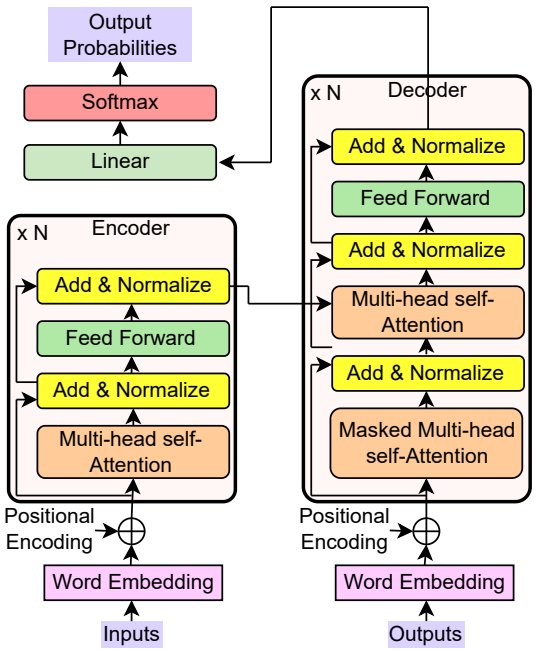


Fig. 1: The `Transformer` architecture proposed by Vaswani [17].

beyond the scope of sequence-to-sequence relations for more general tasks with long-range dependencies, such as the solar irradiance forecast problem considered in this paper.

Furthermore, the self-attention mechanism provides varying weights to each component in the data domain with respect to its relative importance in providing information. In conventional NLP problems, such weighted information is typically referred to as **contexts**. In the solar irradiance problem considered in this paper, such information can be generally considered as the relative importance of aforementioned meteorological factors. The significant improvements introduced by the self-attention mechanism include the following advantages.

- Compared to RNNs, the `Transformer` enables parallel computation to reduce the training time;
- Compared to CNNs, the number of operations required by the `Transformer` to compute the association between two data components does not grow with their distance;
- Attention building blocks can generally avoid the issue of vanishing or exploding gradients.

The operating mechanism of the proposed multi-head `Transformer` is illustrated in Fig. 2. First, each input (e.g., a word in NLP applications or a meteorological factor in solar irradiance applications) is converted into a vector using an embedding algorithm. Secondly, the embedded vector is added by its positional information before it is fed to the (first) encoder (i.e., Encoder 1), which processes all input vectors by two steps:

1) First through a multi-head self-attention layer (the details on multi-head will be discussed in later sections);
2) Then through a feed-forward neural network.

The norm of the summation of the self-attention layer's output and input vectors is calculated between the self-attention layer and the feed forward layer. Finally, the first encoder's output is delivered to the next encoder. Note that

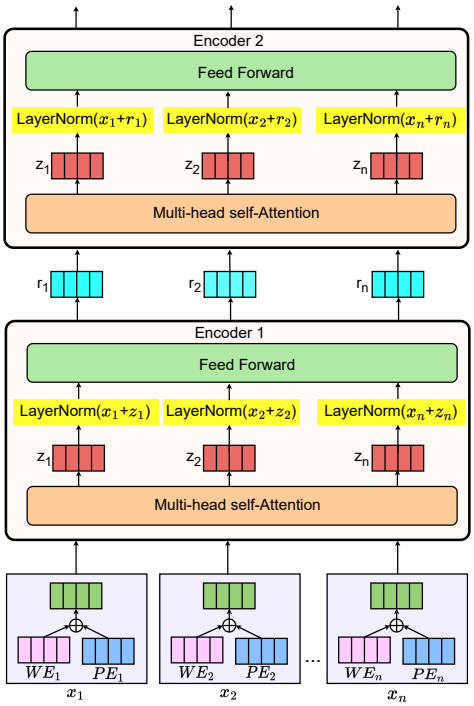


Fig. 2: Performance enhancement via stacking encoders in the `Transformer` architecture.

the stack of encoders (the original `Transformer` is proposed with a stack six encoders) can generally enhance the overall performance. In general the number of stacked encoders depends on data and applications, which is also a trade-off between model complexity and performance.

C. Attention Mechanism

Despite their capabilities in modeling complex structures, RNNs tend to forget previous information, which could cause degraded performances when the length of inputs grows. Moreover, RNNs are often constrained by their computationally inefficiency with respect to sequential inputs. In recent years, the attention mechanism has emerged as a potential solution to these issues. It is a non-uniform weighting method to focus more on one part of the input sequence while giving less attention to the rest to improve the learning process [17]. In RNNs, the output vector p_t of a hidden layer h_t is fed to the attention module as its input. The weight of each hidden state is calculated by

$$p_t = \tanh\left(W^h h_t + b^h\right) \tag{2}$$

$$a_t = \frac{\exp\left(p_t\right)}{\sum_{t=1}^{t_i} \exp\left(p_t\right)} \tag{3}$$

$$h'_t = a_t \odot h_t \tag{4}$$

where α_t and h'_t are the score and attention vectors of h_t, respectively.

1) Self-attention: As discussed above, when the length of input sequences increasingly grows longer, existing deep architectures (e.g., RNNs) tend to "forget" previously learned information, which in turn results in loss of global

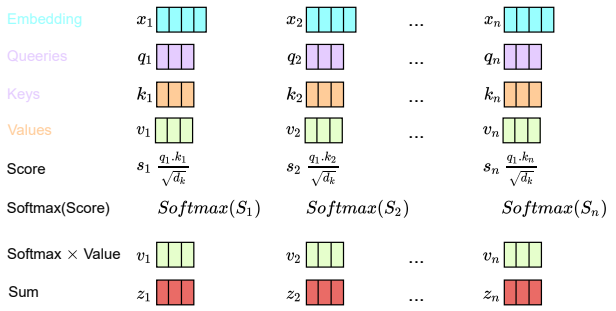


Fig. 3: Single-head self-attention mechanism calculation using vectors [22].

dependencies between input and outputs and consequently degraded learning performance. To overcome this issue, the attention mechanism was introduced in [17] and widely acknowledged, which focuses more on one segment of the input sequence while paying less attention to the others. Self-attention, as a subset of the attention mechanism, is a weighting approach that emphasizes one segment of the input sequence over the others to maximize the learning process. The attention system determines the relationships between two independent sequences, but the self-attention mechanism evaluates the relationships between a sequence and itself. Self-attention is categorized into single-head self-attention and multiple-head self-attention.

2) Single-head Self-attention: The single-head self-attention layer is an agent to find the importance of data at each time in the given query of time-series input data. The mechanism of calculating single-head self-attention is illustrated in Fig. 3. The mechanism of calculating single-head self-attention is illustrated in Figure 3.8. The Q (Query), K (Key) and V (Value) matrices are calculated from input data X. The query vector, Q, is constructed using the current token. The result is compared to the Key vectors, K, of prior tokens. The Value vector, V, contains a representation vector of the current token. The output matrix is calculated by

$$Att(Q, K, V) = \psi\left(\frac{QK^T}{\sqrt{p_k}}\right)V \tag{5}$$

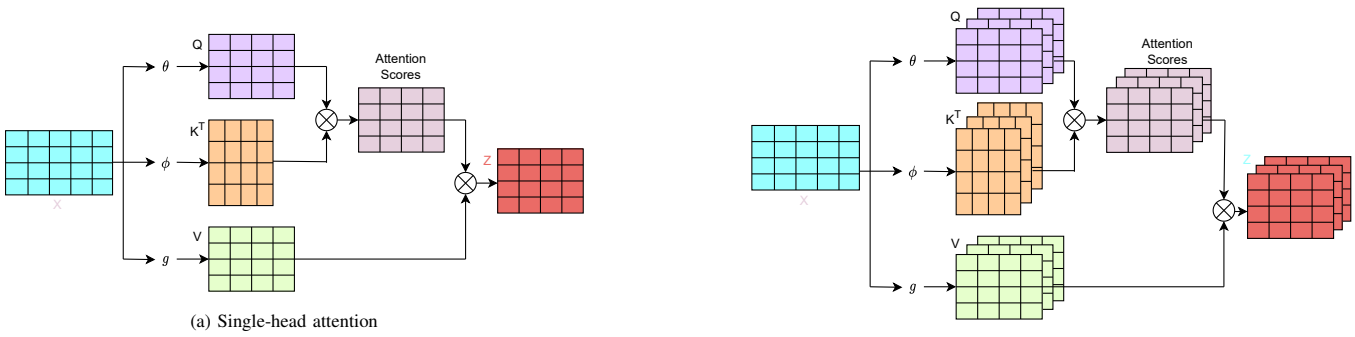
where ψ is the softmax function and p_k is the dimension of K.

3) Multiple-head Self-attention: The aforementioned single-head attention mechanism performs its calculations several times in parallel and concatenates and transforms them non-linearly in the multi-head attention mechanism. Its mathematical representation is illustrated in Figure 4b. It allows the model to learn information simultaneously from different subspaces at various locations. It is formulated as:

$$\text{Multi-head}(Q, K, V) = Concat\left(h_1, \ldots, h_n\right) Z^o \tag{6}$$

$$h_i = Att\left(QZ_i^Q, KZ_i^K, VZ_i^V\right) \tag{7}$$

where h_n donates the n^{th} head and Z^o is the learned projection matrix.



(a) Single-head attention

(b) Multi-head attention

Fig. 4: (a) Mathematical representation for the single-head attention mechanism; and (b) Mathematical representation for the multi-head attention mechanism.

III. Temporal Modeling with Transformer

The original Transformer architecture is optimized for natural processing tasks and may be incompatible with other applications such as time series forecasting. Thus, the self-attentive Transformer architecture is altered for the solar irradiance forecasting task to account for solar irradiance data's variable and stochastic nature. First, the original encoder-decoder Transformer network is optimized for sequence-to-sequence learning, making it ideal for tasks where both input and output are sequences. However, the solar irradiance prediction task takes a sequence as input and return a predicted value as output. Therefore, the proposed architecture has just encoder blocks. Next, the word embedding concept in the NLP task should be changed to time embedding to encode the notion of time in data. It is accomplished by [23], which incorporates an embedding layer into a neural network architecture to learn the time embedding and increase the structure's performance. Finally, the proposed Transformer learns the local variation and pattern of data by using a 1-dimensional convolutional neural network (1D CNN) layer rather than a feed-forward layer, while boosting the generalizability of the system through the use of *dropout*.

A. Time Embedding

As discussed above, in NLP learning tasks, input to the Transformer's encoder blocks are combinations of positional encoding and word embedding, which provides the model with those words' embedding and location in the sentence. Such an emdedding technique has been extended to other applications. For instance, reference [23] incorporates an embedding layer into a neural network architecture in order to learn the time embedding and increase the structure's performance.

To utilize this technique for solar irradiance forecasting, this paper extends word embedding to the time-series forecasting domain by establishing a *time embedding* technique, which can map the temporal correlations between each input sequence. Formally, the time embedding C should be invariant with respect to time and take into account both periodic and non-periodic patterns, which is presented by

$$C(\tau)[j] = \begin{cases} \omega_j \tau + \varphi_j, & \text{if } j = 0 \\ g\left(\omega_j \tau + \varphi_j\right), & \text{if } 0 \leq j \leq k \end{cases} \tag{8}$$

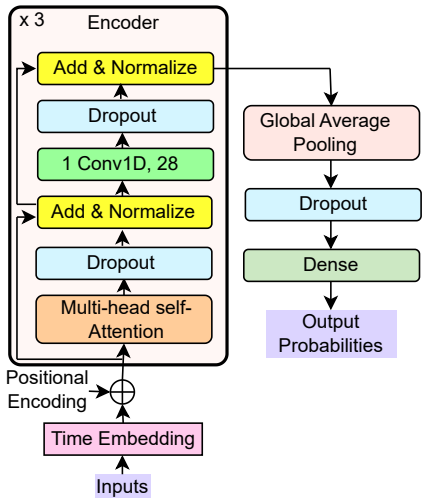


Fig. 5: The proposed Transformer architecture

where $C(\tau)[j]$ donates the j^{th} element of time embedding, g is periodic feature of the time embedding. ω_j, and φ_j are learnable parameters.

B. Proposed Transformer-based Architecture

The proposed architecture is applied independently to each solar site following the data pre-processing stage described in detail in next section III. As illustrated in Figure 5, the Transformer-based temporal model consists of three encoder layers with three heads followed by global average pooling, a dropout, and a dense layer. The embedding of GHI and meteorological values in time is generated and sent to the Transformer encoder layer as input data. The Transformer encoder layer has a multi-head self-attention sub-layer and one 1D CNN sub-layers with 28 kernels with a size and stride of one. Before and after the CNN sub-layer are dropout and normalization layers. The system trains with Adam and MSE optimizer and loss function, receptively for 50 iterations and early stopping with patience 3.

IV. Proposed Spatio-Temporal Architecture Design

The proposed multi-branch hybrid ResNet/Transformer (ResTrans) architecture is shown in Fig. 6, which consists of three major phases. The first phase pre-processes the raw

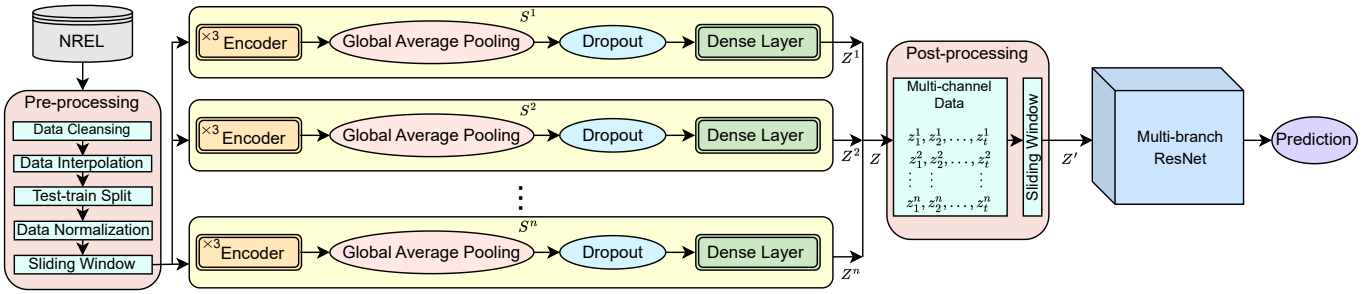


Fig. 6: The proposed spatio-temporal multi-branch ResNet-`Transformer` architecture.

data, which applies cleaning, linear interpolation, and Min-max normalization on every solar site. The second phase is the aforementioned temporal model, which consists of three `Transformer` encoder layers, global average pooling, a dropout, and a dense layer. As illustrated in Figure. 5, the `Transformer`-based temporal model consists of three encoder layers with three heads followed by global average pooling, a dropout, and a dense layer. The embedding of GHI and meteorological values in time is generated and sent to the `Transformer` encoder layer as input data. The `Transformer` encoder layer has a multi-head self-attention sub-layer and one 1D CNN sub-layers with 28 kernels with a size and stride of one. Before and after the CNN sub-layer are dropout and normalization layers.

The proposed architecture is trained with Adam [24] and MSE optimizer and loss function, receptively, for 50 iterations and early stopping with patience 3. The proposed temporal model is applied to each solar site separately. Finally, the outputs from all temporal models are aggregated, rearranged, and fed into the proposed spatial model, a multi-branch ResNet, to provide spatial correlation between solar sites. Note that the proposed mutli-branch ResNet consists of four blocks, including a bottleneck block which is followed by two residual blocks and a prediction block to generate the final estimations. The bottleneck consists of one convolution layer with 64 kernels of size 3 with stride equal to 1, BN layer, and a ReLU layer, with details can be found in [5].

Compared to training multiple architectures separately, the multi-branch architecture computes more efficiently. The multi-branch residual network, learns time-series solar patterns at varying resolutions by cascading multiple convolutional layers with different kernel sizes to analyze solar data at varying resolutions and extract both short and long features along with residual connections to enable the flow of gradient directly through the bottom layers and overcome degradation problems. The output of multi-branch ResNet with J branches and I residual blocks can be calculated by

$$Z_D = Z_0 + \sum_{j=1}^{J} \sum_{i=0}^{I} G\left(Z_{i,j}, \Phi_{i,j}\right) \tag{9}$$

where Z_0 and Z_D donate input and output of ResNet with $Z_{i,j}$ is the input vector of i-th residual block and j-th branch, and $\Phi_{i,j} = \left\{ \phi_{i,j,m} |_{1<m<M} \right\}$ is the set of weights associated with the i-th residual block and j-th branch with M number of layers within residual block. The backpropagation of the total loss function, $\mathcal{L}$, with respect to Z_0 can be defined as

$$\frac{\partial \mathcal{L}}{\partial Z_0} = \frac{\partial \mathcal{L}}{\partial Z_D} \left(\mathbf{1} + \frac{\partial}{\partial Z_0} \sum_{j=1}^{J} \sum_{i=0}^{I} G\left(Z_{i,j}, \Phi_{i,j}\right) \right) \tag{10}$$

where $\mathbf{1}$ denotes the gradient of output directly back-propagates to alleviate the vanishing gradient burden [25].

V. Numerical Validations

A. Data Description

This study utilizes data contains $3,784,320$ observations from the National Solar Radiation Database (NSRDB) [26], which is widely used to examine solar irradiance forecasting performances. The collected dataset consists of 18 sites in Philadelphia, Pennsylvania, from 2000 and 2017 with a 30-minute interval. For performance evaluation, 12 sites are choose in south Philadelphia with 6 additional solar sites around 60 miles from the first 12 sites. All chosen solar sites' geological locations are depicted in Fig. 7.

Fig. 7: Geological locations of 18 selected solar sites used in Philadelphia.

The dataset contains not only the historical GHI $\left(\mathrm{W/m^2}\right)$, DHI $\left(\mathrm{W/m^2}\right)$, and DNI $\left(\mathrm{W/m^2}\right)$ but also the clear-sky GHI $\left(\mathrm{W/m^2}\right)$, clear-sky DHI $\left(\mathrm{W/m^2}\right)$, and clear-sky DNI $\left(\mathrm{W/m^2}\right)$ which represent the maximum values of GHI, DHI, and DNI during clear sky conditions, respectively. Also, NSRDB includes different meteorological measurements such as dew point $(^\circ\mathrm{C})$, solar zenith angle $(^\circ)$, wind speed $(\mathrm{m/s})$, precipitable water (mm), wind direction $(^\circ)$, relative humidity $(\%)$, temperature $(^\circ\mathrm{C})$, pressure (mb), and cloud type. All variables in the dataset are numerical except cloud type, which will be converted to one-hot code. All variables' descriptive statistics are presented in Table I. In general, GHI values increase from 5:00 to 12:00 and then decreasing with time until

TABLE I: Statistics of feature parameters [27].

Variable	Mean	std	Min	50%	Max
Dew Point	7.71	9.52	-24.00	8.00	27.00
Precipitable Water	2.12	1.36	0.10	1.86	7.24
Pressure	1008.20	8.50	940.00	1010.00	1040.00
Relative Humidity	80.98	18.40	18.49	85.71	100.00
Solar Zenith Angle	89.70	36.16	16.45	89.72	163.54
Surface Albedo	0.17	0.20	0.08	0.12	0.87
Temperature	11.70	10.33	-21.00	12.00	39.00
Wind Direction	208.96	99.97	0.00	223.20	360.00
Wind Speed	1.91	1.15	0.00	1.60	10.50
Fill Flag	0.29	0.99	0.00	0.00	4.00
Clearsky DHI	57.36	73.82	0.00	0.00	499.00
Clearsky DNI	305.39	349.83	0.00	0.00	1016.00
Clearsky GHI	229.58	301.86	0.00	0.00	1029.00
DHI	68.00	99.71	0.00	0.00	501.00
DNI	187.66	300.72	0.00	0.00	1016.00
GHI	173.47	257.99	0.00	0.00	1029.00

it drops to zero at around 19:00. Therefore, solar irradiance data between 5:00 and 19:00 is included in the modeling.

Additionally, to explore the presence of spatio-temporal patterns, the correlations between the targeted solar site and its neighbouring solar sites are evaluated. Correlation ranges from -1 to +1. Closer to zero values indicate that there is no linear relationship between the two variables. The closer the correlation is to 1, the more positively correlated it is.

B. Data Process

The dataset is split into a training set (consisting of data from years 2000–2010), a validation set (2011–2013), and a test set (2014–2017). All observations with missing data and superfluous data are interpolated with a linear interpolation technique. The numerical values are normalized by Min-Max and converted into a normalized number between 0 and 1:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{11}$$

where x is the original value and x' is its normalized value. The sliding window method generates data points, $X_t'^n$. The sliding window technique generates data for the current time step by moving a window of a specified length over the historical data, sample by sample. If the data of n-th solar site is presented by matrix $X_t^n \in \mathbb{R}^K, K$ is number of variables, sliding window generates $X_t'^n \in \mathbb{R}^{K\times L}, L$ is window length. In this study, the same sliding window length is applied to all input variables and is set to 24, representing 24 time steps for the considered half-day solar irradiance forecasting and the observation interval of 30 minutes.

C. Performance Metrics

Typically, researchers report the performance of solar forecasting systems in terms of MAE, RMSE, and R^2 coefficient [28], [29]. The RMSE is a quadratic scoring rule that determines the average magnitude of the error. RMSE is the most reliable evaluation metrics since it assists in detecting and removing the outliers in the data. It is sensitive to large forecasting errors while being forgiving of minor errors

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(z_t - \hat{z}_t)^2}, \tag{12}$$

Algorithm 1: Multi-Branch ResTran

Input $X \in \mathbb{R}^{N\times K\times T}$, learning rates α and γ
`// Data Pre-processing`
for $S \leftarrow 1, 2, \ldots, n$ **do**
  Data cleansing, Interpolation of missing data
  Data normalization
  Dividing data to training, validation, and test sets
  Data Embedding
end
return X'
`// Training transformer with Attention`
for $S \leftarrow 1, 2, \ldots, n$ **do**
  Input $\leftarrow X'$, Initialize $\{Q, K, V\}$
  while *Stopping criterion not meet* **do**
    Multi-head self-attention, parallel calculation
    Calculate the total lost function using MSE
    Accelerate with Adaptive moment estimation
    Output sequence, $\hat{Z}_t^n$
  end
end
return Predictions, $\hat{Z}^n \in \mathbb{R}^T$
return Predictions of all sites, $\hat{Z} \in \mathbb{R}^{N\times T}$
return Sliding window data Z'
`// Training Multi-Branch ResNet`
Input $\leftarrow Z'$, Initialize $\{\Phi\}$
while *Stopping criterion not meet* **do**
  Sample from the training set
  Calculate prediction, $\hat{Y}_t \in \mathbb{R}^N$
  Calculate the total lost function using MSE
  Compute gradient estimate, β
  Update weights $\Phi \leftarrow \Phi - \gamma\beta$
end
return Final predictions, $\hat{Y} \in \mathbb{R}^{N\times T}$
Output $Y \in \mathbb{R}^{N\times T}$

where z_t and $\hat{z}_t$ are the actual and predicted GHI at target solar site, respectively.

The MAE is a linear score, meaning that all individual differences are equally weighted on average. Hence, it provides equal weight to all differences in the data.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|z_t - \hat{z}_t|, \tag{13}$$

A lower value indicates a more accurate forecasting estimation.

The R^2 is a statistical measure that that indicates how much the prediction model deviates from reality.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(z_t - \hat{z}_t)^2}{\sum_{t=1}^{n}(\bar{z} - \hat{z}_t)^2}, \tag{14}$$

where $\bar{z}$ donates the mean of the actual data. The R^2 output is a number between 0 and 1, where 0 indicates that the model fits poorly and 1 indicates that the model fits perfectly.

VI. Numerical Results and Analysis

This section presents results that justify at the proposed hybrid ResTran architecture's efficiency for geographically distributed solar sites.

1) Transformer as a Temporal Model: This section presents numerical evaluation results and compares the Transformer-based temporal architecture's performance with benchmark models, including CNN, LSTM, AttLSTM (proposed by earlier work [5]), and ResNet.

Table II compares the performance of benchmark models with self-attentive Transformer architecture across a 12-hour time horizon. The average RMSE, MAE, and R^2 of baseline models across the 12 solar sites are as follows: CNN (0.19, 0.13, 0.63), LSTM (0.14, 0.09, 0.77), AttLSTM (0.12, 0.07, 0.78), and ResNet (0.08, 0.05, 0.84). Furthermore, recurrent algorithms outperform CNN because of feedback loops that allow them to recall information from the past and learn short-term and long-term dependencies. The average RMSE and MAE of CNN, respectively, are 0.19 and 0.13, and that of the LSTM is 0.14 and 0.09, which shows 26.31%, 30.77% of improvement.

It can be further observed in Table II that integrating LSTM with an attention mechanism, AttLSTM improves LSTM performance by 14.28%, 22.22%, and 1.27% in terms of RMSE, MAE, and R^2, respectively. Moreover, as the LSTM output is reliant on the output of prior states at each time step, this might result in a forgetting problem and performance degradation in an excessively lengthy input sequence. This problem is solved in the Transformer by processing the input sequence as a whole, utilizing the self-attention mechanism, and avoiding recursion.

The Transformer, with an average RMSE of 0.06 W/m^2, an average MAE of 0.04 W/m^2, and a R^2 coefficient of 0.87%, beats LSTM by 57.14%, 55.56%, and 12.99% in terms of RMSE, MAE, and R^2, respectively. As seen in III, both ResNet and Transformer are deep architectures with 985,361 and 1,106,011 parameters, respectively. However, Transformer training time is 17.35% faster than ResNet, despite having 12.24% more parameters. This is because the Transformer employs a parallel training process, whereas the ResNet uses a sequential procedure. Furthermore, the Transformer surpasses ResNet by 25.00% and 20.00% in terms of average RMSE and MAE, respectively. Thus, we can conclude that the self-attentive Transformer architecture introduced here considerably improve forecasting accuracy and surpasses all other deep learning baseline models.

A. Model efficiency on nearby solar sites on forecast errors

The suggested multi-branch ResTrans architecture is compared to baseline deep learning models such as CNN-LSTM [30], single-branch ResLSTM [30], and multi-branch ResLSTM [5]. Finally, the superiority of the multi-brach ResTrans model is established by comparison to the single-branch ResTrans architecture. Table IV compares the performance of benchmark models with the proposed multi-branch ResTrans in terms of average RMSE, average MAE, and R^2 across the 12 solar sites in south Philadelphia. By substituting LSTMs with the Transformer in single-branch and multi-branch ResLSTM architectures, the average RMSE of the architectures was decreased by 8.70% and 14.04%, respectively. Compared to LSTM, which has issues such as vanishing gradients and forgetting earlier data, the Transformer maintains direct connections to all preceding timestamps, allowing information propagation over considerably longer sequences.

Moreover, comparing the average RMSE and MAE of single-branch ResTrans and multi-branch ResTrans architectures demonstrates that multi-branch ResTrans improves the forecasting accuracy by 22.22% and 20.51% in terms of average RMSE, MAE, respectively. It is due to the fact that solar irradiance data has various embedded information at different resolutions, and single resolution architectures extract features at one resolution while ignoring features at others. Therefore, the multi-branch network overcomes this limitation by modeling data at various resolutions and concurrently learning local and global trends to improve prediction accuracy [5]. The numerical results prove that the proposed multi-branch ResTrans model is good at capturing the uncertainty and non-linearity of solar irradiance.

Moreover, an illustrative plot of the actual GHI and forecasted GHI by both the benchmarks and by the proposed multi-branch ResTrans are shown in Fig. 8. It can be observed in Fig. 8 that the proposed multi-branch ResTrans outperforms benchmark models in term of closely matching the actual GHI measurements, which aligns with the numerical results shown in Table II.

B. Model efficiency on distant solar sites on forecast errors

The performance of the proposed spatio-temporal architectures on all 18 locations is summarized in Table V. Furthermore, Tables IV and V indicate that when features from distant solar sites are included in the study, the models' accuracy slightly decrease due to their lower correlation with distant solar sites locations. For instance, LSTM-CNN delivers the poorest results, with an average RMSE of 0.078 (W/m^2), an average MAE of 0.053 (W/m^2), and a R^2 coefficient of 0.79, corresponding to a 1.3% and 1.9% increase in RMSE and MAE, respectively. The single-branch and Multi-branch ResLSTM designs yielded a higher average RMSE of 2.9% and 5.3%, respectively, and a higher average MAE of 4.08% and 7.9%, respectively, compared to the similar structures in Table IV. Furthermore, the average RMSE and MAE of the single-branch ResTrans architecture are 0.067 (W/m^2) and 0.046 (W/m^2), respectively, which are 6.3% and 17.9% greater than the similar structure in Table IV.

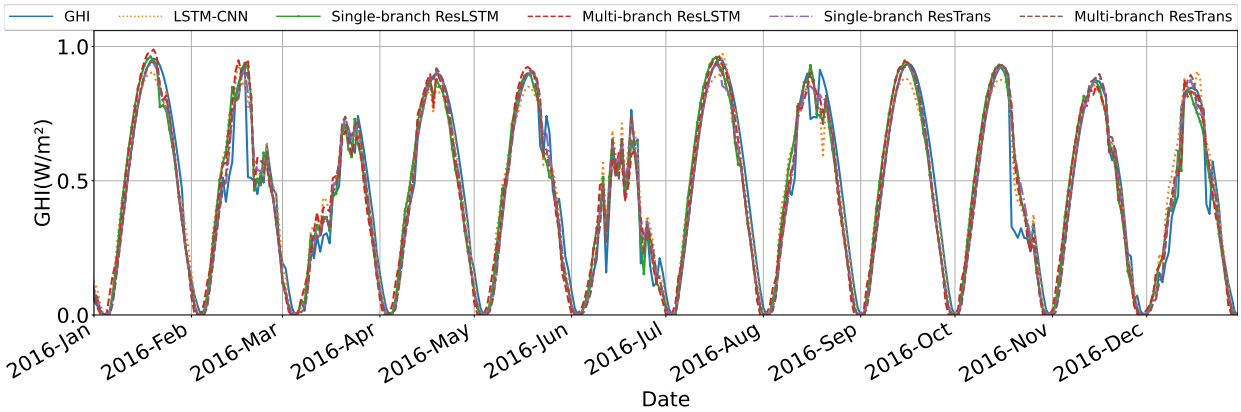
The findings demonstrate that there is a negative correlation between forecasting accuracy and distance.

VII. Conclusion

This paper introduces a novel spatio-temporal multi-resolution ResTrans network for short-term solar irradiation forecasting. The proposed hybrid model aimed at separating temporal models from spatial models to alleviate the curse of dimensionality. First, the temporal model is applied to

TABLE II: Evaluating temporal `Transformer` architecture in comparison to baseline temporal models

Solar Sites	RMSE (W/m^2)					MAE (W/m^2)					R^2				
	CNN	LSTM	AttLSTM	SB ResNet	Trans.	CNN	LSTM	AttLSTM	SB ResNet	Trans.	CNN	LSTM	AttLSTM	SB ResNet	Trans.
#1	0.17	0.11	0.09	0.07	**0.06**	0.11	0.07	0.05	0.06	**0.04**	0.69	0.83	0.84	0.84	**0.87**
#2	0.19	0.12	0.11	0.08	**0.08**	0.12	0.08	0.06	0.05	**0.05**	0.63	0.79	0.80	0.83	**0.84**
#3	0.19	0.15	0.13	0.09	**0.07**	0.12	0.10	0.11	0.06	**0.04**	0.63	0.75	0.77	0.85	**0.86**
#4	0.19	0.13	0.11	0.08	**0.05**	0.13	0.08	0.06	0.05	**0.03**	0.60	0.78	0.77	0.85	**0.88**
#5	0.18	0.13	0.12	0.09	**0.06**	0.12	0.09	0.05	0.06	**0.04**	0.65	0.76	0.78	0.87	**0.88**
#6	0.19	0.13	0.10	0.09	**0.05**	0.12	0.09	0.06	0.06	**0.03**	0.60	0.76	0.79	0.87	**0.88**
#7	0.19	0.16	0.13	0.09	**0.07**	0.13	0.10	0.07	0.05	**0.04**	0.69	0.75	0.76	0.84	**0.85**
#8	0.22	0.12	0.09	0.09	**0.06**	0.15	0.08	0.06	0.05	**0.04**	0.56	0.79	0.80	0.86	**0.87**
#9	0.20	0.12	0.11	0.08	**0.05**	0.13	0.08	0.06	0.05	**0.03**	0.67	0.79	0.81	0.85	**0.88**
#10	0.19	0.14	0.12	0.08	**0.07**	0.13	0.10	0.09	0.05	**0.04**	0.63	0.73	0.75	0.84	**0.86**
#11	0.18	0.14	0.12	0.08	**0.05**	0.12	0.10	0.08	0.05	**0.03**	0.66	0.72	0.75	0.83	**0.89**
#12	0.19	0.16	0.15	0.08	**0.07**	0.13	0.11	0.10	0.05	**0.05**	0.60	0.75	0.76	0.82	**0.85**
Avg.	0.19	0.14	0.12	0.08	**0.06**	0.13	0.09	0.07	0.05	**0.04**	0.63	0.77	0.78	0.84	**0.87**



(a) Illustration of actual GHI vs. forecasted GHI by both the benchmarks and by the proposed multi-branch ResTrans over a year (2016) of processed test data. Each day's GHI is only shown between 5 AM to 7 PM and represented by 28 data points (one point every 30 minutes between 5 AM to 7 PM).

Fig. 8

TABLE III: Comparison of the number of parameters, training time

Models	# Parameters	Training Time (s)
Single-branch ResNet	985,361	15,465
`Transformer`	1,106,011	12,782

each solar site separately to learn the temporal pattern and perform dimension reduction. Then, outputs of the temporal modeling step are fed to the proposed ResTran for spatial modeling. The proposed hybrid ResTran architecture inherits `Transformer`'s designed capability of effectively handling lengthy input sequences in parallel as well as multi-resolution ResNet's strength in learning local and global patterns. Consequently, the proposed ResTran can accurately and effectively model historical solar irradiance in a spatio-temporal manner, which is also validated by the numerical results. Finally, the proposed ResTran's effectiveness against distant solar sites was also investigated. The numerical results indicated that the accuracy of forecasts reduces as the distance between solar sites grows.

REFERENCES

[1] S. Ziyabari, L. Du, and S. Biswas, "Multi-branch resnet-transformer based deep hybrid approach for short-term spatio-temporal solar irradiance forecasting," in *2022 IEEE Energy Conversion Congress and Exposition (ECCE)*, 2022, pp. 1–5.

[2] U.S. Federal Energy Regulatory Commission, "Order No. 2222: Participation of Distributed Energy Resource Aggregations in Markets Operated by Regional Transmission Organizations and Independent System Operators," issued September 17, 2020.

[3] X. Fan, D. Moscovitz, L. Du, and W. Saad, "A data-driven democratized control architecture for regional transmission operators," in *2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2022, pp. 1–5.

[4] Z. Zhao, D. Moscovitz, S. Wang, X. Fan, and L. Du, "Semi-supervised disaggregation of load profiles at transmission buses with significant behind-the-meter solar generations," in *2022 IEEE Energy Conversion Congress and Exposition (ECCE)*, 2022, pp. 1–5.

[5] S. Ziyabari, L. Du, and S. Biswas, "Multibranch attentive gated resnet for short-term spatio-temporal solar irradiance forecasting," *IEEE Trans. Ind. Appli.*, vol. 58, no. 1, pp. 28–38, 2022.

[6] A. Dolatabadi, H. H. Abdeltawab, and Y. A.-R. I. Mohamed, "Deep reinforcement learning-based self-scheduling strategy for a caes-pv system using accurate sky images-based forecasting," *IEEE Transactions on Power Systems*, 2022.

[7] A. Asrari, T. X. Wu, and B. Ramos, "A hybrid algorithm for short-term solar power prediction—sunshine state case study," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 582–591, 2016.

[8] Z. Zhen *et al.*, "Pattern classification and pso optimal weights based sky images cloud motion speed calculation for solar pv power forecasting," *IEEE Trans, Ind. Appl.*, vol. 55, no. 4, pp. 3331–3342, 2019.

[9] Z. Si, Y. Yu, M. Yang, and P. Li, "Hybrid solar forecasting method using satellite visible images and modified convolutional neural networks," *IEEE Trans, Ind. Appl.*, vol. 57, no. 1, pp. 5–16, 2021.

[10] R. Zhang, H. Ma, W. Hua, T. K. Saha, and X. Zhou, "Data-Driven Photovoltaic Generation Forecasting Based on a Bayesian Network with Spatial-Temporal Correlation Analysis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, 2020.

[11] M. Khodayar, S. Mohammadi, M. E. Khodayar, J. Wang, and G. Liu, "Convolutional graph autoencoder: A generative deep neural network for probabilistic spatio-temporal solar irradiance forecasting," *IEEE Trans. Sustain. Energy*, vol. 11, no. 2, pp. 571–583, 2020.

TABLE IV: Comparison of numerical evaluations and performances of spatio-temporal architectures on 12 solar sites

Solar Sites	RMSE (W/m^2)					MAE (W/m^2)					R^2				
	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans
#1	0.076	0.067	0.053	0.063	**0.044**	0.052	0.048	0.037	0.040	**0.029**	0.82	0.89	0.96	0.93	**0.98**
#2	0.070	0.062	0.057	0.059	**0.047**	0.049	0.045	0.038	0.041	**0.031**	0.83	0.90	0.96	0.94	**0.98**
#3	0.080	0.070	0.054	0.060	**0.044**	0.054	0.049	0.037	0.041	**0.029**	0.81	0.88	0.96	0.91	**0.97**
#4	0.066	0.071	0.057	0.062	**0.051**	0.046	0.049	0.038	0.042	**0.032**	0.83	0.88	0.95	0.91	**0.96**
#5	0.072	0.064	0.053	0.060	**0.044**	0.050	0.046	0.037	0.041	**0.029**	0.82	0.86	**0.98**	0.90	**0.98**
#6	0.078	0.068	0.059	0.063	**0.052**	0.053	0.049	0.040	0.043	**0.033**	0.80	0.87	0.95	0.90	**0.97**
#7	0.091	0.074	0.060	0.064	**0.052**	0.059	0.051	0.040	0.045	**0.033**	0.79	0.88	0.95	0.91	**0.98**
#8	0.081	0.069	0.053	0.064	**0.044**	0.055	0.048	0.036	0.045	**0.030**	0.81	0.89	0.96	0.91	**0.98**
#9	0.075	0.059	0.057	0.058	**0.047**	0.052	0.044	0.038	0.039	**0.031**	0.79	0.90	0.96	0.93	**0.97**
#10	0.070	0.071	0.057	0.065	**0.051**	0.048	0.049	0.038	0.045	**0.032**	0.82	0.87	0.95	0.91	**0.95**
#11	0.080	0.077	0.060	0.068	**0.054**	0.053	0.052	0.040	0.046	**0.034**	0.79	0.84	0.95	0.88	**0.96**
#12	0.080	0.073	0.060	0.066	**0.052**	0.054	0.051	0.040	0.044	**0.033**	0.79	0.88	0.95	0.90	**0.97**
Avg.	0.077	0.069	0.057	0.063	**0.049**	0.052	0.049	0.038	0.039	**0.031**	0.81	0.88	0.96	0.90	**0.97**

TABLE V: Comparison of numerical evaluations and performances of spatio-temporal architectures on 18 solar sites

Solar Sites	RMSE (W/m^2)					MAE (W/m^2)					R^2				
	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans	LSTM-CNN	Single-br. ResLSTM	Multi-br. ResLSTM	Single-br. ResTrans	Multi-br. ResTrans
#1	0.076	0.73	0.056	0.065	**0.047**	0.053	0.050	0.043	0.042	**0.031**	0.80	0.86	0.93	0.91	**0.95**
#2	0.079	0.066	0.060	0.061	**0.048**	0.051	0.048	0.042	0.045	**0.033**	0.82	0.87	0.92	0.93	**0.94**
#3	0.078	0.074	0.057	0.063	**0.046**	0.057	0.051	0.042	0.046	**0.032**	0.77	0.85	0.94	0.90	**0.95**
#4	0.065	0.075	0.059	0.064	**0.053**	0.048	0.053	0.044	0.044	**0.035**	0.80	0.87	0.92	0.89	**0.94**
#5	0.074	0.066	0.056	0.066	**0.045**	0.053	0.049	0.041	0.042	**0.030**	0.82	0.83	0.94	0.90	**0.95**
#6	0.077	0.071	0.059	0.067	**0.054**	0.055	0.054	0.045	0.047	**0.034**	0.78	0.85	0.92	0.91	**0.94**
#7	0.093	0.078	0.062	0.066	**0.055**	0.061	0.055	0.045	0.049	**0.035**	0.78	0.86	0.93	0.90	**0.95**
#8	0.085	0.072	0.056	0.063	**0.047**	0.057	0.050	0.041	0.050	**0.034**	0.79	0.87	0.94	0.89	**0.95**
#9	0.077	0.063	0.058	0.062	**0.048**	0.054	0.049	0.042	0.045	**0.036**	0.77	0.86	0.94	0.92	**0.95**
#10	0.070	0.075	0.058	0.067	**0.054**	0.050	0.052	0.042	0.048	**0.035**	0.80	0.83	0.90	0.90	**0.92**
#11	0.080	0.080	0.063	0.069	**0.056**	0.056	0.055	0.044	0.051	**0.036**	0.80	0.81	0.89	0.88	**0.91**
#12	0.081	0.076	0.061	0.070	**0.055**	0.055	0.057	0.042	0.048	**0.037**	0.79	0.87	0.92	089	**0.93**
#13	0.075	0.065	0.063	0.070	**0.056**	0.057	0.049	0.041	0.039	**0.038**	0.75	0.82	0.91	0.85	**0.94**
#14	0.077	0.072	0.058	0.069	**0.049**	0.051	0.050	0.041	0.047	**0.035**	0.76	0.83	0.90	0.86	**0.93**
#15	0.076	0.069	0.061	0.071	**0.048**	0.050	0.048	0.043	0.044	**0.032**	0.78	0.86	0.91	0.89	**0.92**
#16	0.081	0.070	0.059	0.72	**0.046**	0.049	0.053	0.044	0.049	**0.034**	0.77	0.87	0.91	0.90	**0.93**
#17	0.079	0.068	0.065	0.069	**0.044**	0.050	0.054	0.041	0.046	**0.034**	0.80	0.85	0.92	0.88	**0.92**
#18	0.078	0.071	0.065	0.069	**0.046**	0.053	0.049	0.040	0.044	**0.031**	0.78	0.86	0.90	0.91	**0.92**
Avg.	0.078	0.071	0.060	0.067	**0.050**	0.053	0.051	0.042	0.046	**0.034**	0.79	0.85	0.92	0.89	**0.94**

[12] S. Chai, Z. Xu, Y. Jia, and W. K. Wong, "A robust spatiotemporal forecasting framework for photovoltaic generation," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5370–5382, 2020.

[13] P. Kumari and D. Toshniwal, "Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting," *Appl. Ener.*, vol. 295, p. 117061, 2021.

[14] J. Simeunović, B. Schubnel, P.-J. Alet, and R. E. Carrillo, "Spatio-temporal graph neural networks for multi-site pv power forecasting," *IEEE Trans. Sustain. Energy*, vol. 13, no. 2, pp. 1210–1220, 2022.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[16] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3285–3292.

[17] A. Vaswani *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[18] Y. Li, X. Peng, J. Zhang, Z. Li, and M. Wen, "Dct-gan: Dilated convolutional transformer-based gan for time series anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[19] Y. Wu, C. Lian, Z. Zeng, B. Xu, and Y. Su, "An aggregated convolutional transformer based on slices and channels for multivariate time series classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–12, 2022.

[20] G. Harerimana, J. W. Kim, and B. Jang, "A multi-headed transformer approach for predicting the patient's clinical time-series variables from charted vital signs," *IEEE Access*, vol. 10, pp. 105 993–106 004, 2022.

[21] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[22] S. Ziyabari, L. Du, and S. K. Biswas, "Short-term solar irradiance forecasting based on self-attentive transformers," in *2022 IEEE Power & Energy Society General Meeting (PESGM)*, 2022, pp. 1–5.

[23] S. M. Kazemi *et al.*, "Time2vec: Learning a vector representation of time," *arXiv preprint arXiv:1907.05321*, 2019.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," *Renewable and sustainable energy reviews*, vol. 89, pp. 51–60, 2018.

[27] S. S. Khoshgoftar Ziyabari, "Short-term spatio-temporal solar irradiance forecasting using multi-resolution deep learning models," Ph.D. dissertation, Temple University, 2022.

[28] H. Zhou *et al.*, "Short-Term photovoltaic power forecasting based on long short term memory neural network and attention mechanism," *IEEE Access*, vol. 7, pp. 78 063–78 074, 2019.

[29] M. Abdel-Nasser, K. Mahmoud, and M. Lehtonen, "Reliable solar irradiance forecasting based on Choquet integral and deep LSTM," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1873–1881, 2021.

[30] S. Ziyabari, L. Du, and S. Biswas, "Short-term solar irradiance forecasting based on multi-branch residual network," in *2020 IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 2000–2005.

Saeedeh Ziyabari received her B.Eng. degree in computer engineering from Azad University South Tehran Branch, Tehran, Iran, Master's degree in smart technology and robotic engineering from University Putra Malaysia (UPM), Malaysia, and Ph.D. degree in electrical and computer engineering from Temple University in 2002, 2017, and 2022, respectively. Her research interests include solar energy forecasting techniques using advanced machine learning.

Zhenyu Zhao (S'23) received his B.Eng degree in Automation from Wuhan University of Technology, China in 2018 and his M.S. degree in Electrical Engineering from the George Washington University, Washington, DC in 2020. He is currently pursuing his Ph.D. degree at Temple University, Philadelphia, PA. He is an intern at PJM Interconnection (Audubon, PA) for summer 2023. His research interests include advanced data analytical and learning methods with applications to transmission systems.

Liang Du (S'09–M'13–SM'18) received the Ph.D. degree in electrical engineering from Georgia Institute of Technology, Atlanta, GA in 2013. He was a Research Intern at Eaton Corp. Innovation Center (Milwaukee, WI), Mitsubishi Electric Research Labs (Cambridge, MA), and Philips Research N.A. (Briarcliff Manor, NY) in 2011, 2012, and 2013, respectively. He was an Electrical Engineer with Schlumberger, Sugar Land, TX, from 2013 to 2017. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering at Temple University, Philadelphia.

Dr. Du received the Ralph E. Powe Junior Faculty Enhancement Award from ORAU in 2018, Early-Career Fellowship from National Academies of Science, Engineering, and Medicine (NASEM) in 2022, and CAREER award from National Science Foundation in 2023. He currently serve as an associate editor of IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS, IEEE TRANSACTIONS ON SUSTAINABLE ENERGY, and IEEE TRANSACTIONS ON TRANSPORTATION ELECTRIFICATION, Program Co-Chair and Treasurer of 2023 IEEE Transportation Electrification Conf. & Expo. (ITEC23), as well as the Mentorship Activity Chair of IEEE Power Electronics Society (PELS) Technical Committee 4 (TC-4).

Saroj K. Biswas (S'85–M'85-LSM'21) is a Professor of Electrical and Computer Engineering at Temple University, Philadelphia, specializing in control and optimization of dynamic systems, multiagent systems, power systems, and distributed parameter systems. Dr. Biswas received his Ph.D. in Electrical Engineering from University of Ottawa, Canada in 1985, and master's and bachelor's degrees in Electrical Engineering from Bangladesh University of Engineering and Technology in 1977 and 1975, respectively. He is an Associate Editor of *Journal of Industrial and Management Optimization* (JIMO), and a former Associate editor of *Dynamics of Continuous, Discrete and Impulsive Systems* (DCDIS-B). Dr. Biswas is a Life Senior Member of IEEE, and member of ASEE, and Sigma Xi.