TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2026.9010026

Ns_Transraph: Transformer-based spatio-temporal aggregation framework with implicit graph modeling for photovoltaic forecasting

Ke Wang, Shibo Wang, Ji Wu, Jingxin Zhang*

Abstract: Accurate photovoltaic (PV) forecasting is extremely important for power grid dispatching and power security. Generally, the power forecasting for a single PV station relies on its corresponding data, which neglects the valuable information of adjacent stations. Besides, the nonstationarity of PV power has not been considered yet. Against this background, this paper investigates a novel ultra-short-term PV forecasting approach called ns_Transraph (**ns_Trans**former g**raph** modeling), which could utilize data from the adjacent stations to enhance the forecasting performance. Specifically, transfer entropy is used to select adjacent stations that provide beneficial information to the target station. Then, an efficient method called Transraph is proposed based on the Transformer and implicit graph modeling, to extract the spatio-temporal relationship of adjacent PV stations simultaneously. Moreover, a cross-attention aggregation mechanism is designed to aggregate the information adaptively within the graph structure. To deal with nonstationarity, a de-stationary module is embedded in the attention layer to focus on the target station and aggregate the data of adjacent PV stations simultaneously. Experimental results illustrate that the proposed ns_Transraph method outperforms several state-of-the-art methods in the medium and long-term prediction on both practical datasets.

Key words: Ultra-short-term photovoltaic forecasting, cross attention aggregation mechanism, spatio-temporal information, Transformer, implicit graph model

1 Introduction

To satisfy the increasing demand for clean and sustainable energy, PV energy receives increasing attention and has been widely utilized throughout the world. However, the PV station is greatly affected by complex weather conditions, which make it uncertain and unstable. Accurate forecasting of ultra-short-term PV generation is crucial for grid dispatching and grid stability. Therefore, relevant research has yielded fruitful achievements.

PV forecasting methodologies can be categorized into physical and statistical methods. The physical methods, based on numerical weather forecasting, employ physical models and choose physical data to generate forecasts[1]. Ensembling numerical weather forecasting and physical models is adopted to obtain excellent forecasting[2]. This methodology is suitable for newly built PV stations, as it avoids the dependency on sufficient historical data.

The statistical methods extract the pattern and correlation from past data[1]. Statistical methods can be further categorized into machine learning methods and deep learning (DL) methods. In recent years, the superiority of DL methods in time series forecasting has attracted attention[3, 4, 5], and several studies have illustrated their promising potential in

• Ke Wang is with Southeast University-Monash University Joint Graduate School, Southeast University, Suzhou, China. Email: 220245123@seu.edu.cn.
• Shibo Wang is with the Electric Power Research Institute, State Grid Shandong Electric Power Company, Jinan, China. Email: wangshibodu@126.com.
• Ji Wu is with the China Electric Power Research Institute, Nanjing, China. Email: wuji419520373@126.com
• Jingxin Zhang is with the School of Automation, and also with Institute of Intelligent Unmanned Systems, Southeast University, Nanjing, China. Email: jingxinzhang@seu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2025-05-29; revised: 2025-10-03; accepted: 2026-02-10

photovoltaic power forecasting(PVPF). For instance, Srivastava et al. utilized long short-term memory (LSTM) networks and concluded that deep neural networks offered significant capabilities for energy modeling[6]. To enhance forecasting performance, some researchers have explored hybrid approaches by combining different DL models. Wang et al. developed a hybrid LSTM-convolutional neural network (CNN) architecture that leveraged the strengths of both models[7]. Similarly, Yin et al. proposed a framework involving multiple groups of well-trained regression networks with varying configurations, which were subsequently integrated using a weighted fully connected regression network[8]. Furthermore, Liu et al. introduced an integrated model combining parallel bidirectional LSTM (BiLSTM) and CNN architectures. The model employs multimodal decomposition to boost forecasting accuracy[9]. However, these DL models typically require large volumes of data for effective training. Currently, models relying solely on single station data exhibit diminishing returns, indicating a potential performance bottleneck.

In recent decades, the increasing number of PV stations has spurred a growing body of research on PVPF employing data from adjacent PV stations. Incorporating data from adjacent PV stations can enhance forecasting performance owing to abundant features. Consequently, many studies have focused on methods integrating data from adjacent PV stations to improve forecasting accuracy. Spatial and temporal aggregation is becoming the main research direction[10]. Wen et al.[11] enhanced the FEDformer model by adding multiple multi-layer perceptron layers, thereby improving its ability to capture deep features. They also leveraged the known geographic location information of PV stations by embedding it directly into the decoder's input. Similarly, Lai et al.[12] employed a statistical upscaling method to estimate regional power output. They classified adjacent PV stations into different regions and developed separate forecasting models for each region.

Spatial information also makes sense according to the improvements researchers have achieved in recent decades. Some researchers applied graph neural network to extract the spatial features between adjacent PV stations[13]. Song et al.[14] proposed a framework to forecast PV generation of large-scale stations and deal with spatio-temporal missing data. Bai et al.[15] combined the advantages of graph convolutional network (GCN) in spatial feature extraction and the advantages of gated recurrent unit (GRU) network in temporal feature extraction. Zhang et al.[16] proposed a model based on the dynamic graph convolutional layer and BiLSTM module. Yang et al.[17] proposed DEST-GNN based on GCN and temporal convolutional networks (TCN), which extracted spatial and temporal information using the sparse attention block. Gao et al.[18] introduced a spatiotemporal hybrid model that embedded Chebyshev graph convolution into BiLSTM to capture spatial and temporal dependencies simultaneously. By leveraging multi-graph structures and attention-driven fusion, the model effectively enhances both single-station and multi-station photovoltaic power prediction accuracy. Bai et al.[19] constructed adjacency matrices using the maximum information coefficient to extract spatial features from multiple stations and attribute spatial information with GRU. Although the above methods improve forecasting accuracy, several challenges remain unresolved:

(1) Most existing approaches model spatial and temporal dependencies separately. However, under diverse and rapidly changing weather conditions, the interactions between spatial and temporal dependencies across PV stations are highly coupled. Modeling them independently makes it difficult to capture complex spatio-temporal interactions.

(2) As the depth of graph-based aggregation increases, models tend to suffer from over-smoothing. And the representation of the target station is excessively influenced by neighboring stations. In addition, the inherent non-stationarity of PV power generation at the target station is often insufficiently preserved.

To overcome the aforementioned problems, ns_Transraph is proposed based on the Transformer and graph modeling structures. A cross-attention aggregation mechanism is introduced in the encoder to improve the information aggregation of adjacent stations through implicit graph modeling. Moreover, de-stationary cross attention is introduced to focus on the nonstationarity of the original target sequence and alleviate the over-smoothing caused by information aggregation. The main contributions of this work are summarized as follows:

(1) A new method is investigated based on Transformer and implicit graph modeling together, enabling the model to extract global information. The graph structure is integrated into the encoder part of the Transformer. To the best of our knowledge, the cross-attention mechanism is the first used to aggregate the information between nodes of the graph structure.

(2) The multidimensional data is processed by Transformer and implicit graph structure in the encoder. This can avoid time series data splicing and excessive stacking of graph layers, which preserves the time feature and the differentiation of nodes. Spatio-temporal relationships between the target station and the adjacent stations are extracted simultaneously, preserving the complex spatial and temporal dependencies.

(3) Aiming at the nonstationary features of the target station, the de-stationary cross-attention mechanism is incorporated to avoid the over-smoothing and excessive stationarity simultaneously.

The remainder of this article is structured as follows. Section 2 reviews the basic theory of three relevant algorithms and formulates the problem. Section 3 outlines the proposed methodology, including overall architecture, the cross attention aggregation mechanism and the de-stationary cross attention. Section 4 presents the experimental results along with an analysis of the findings. Finally, Section 5 concludes this paper and discusses potential directions for future research.

2 Preliminaries and problem statement

In this section, the related studies of transfer entropy analysis, graph-structured modeling and transformer are described briefly, and the problem statement is formulated thereafter.

2.1 Transfer entropy analysis

Given the complex nonlinear relationships among PV stations, transfer entropy[20] serves as an effective method to quantify the dependencies between a target station and its adjacent stations. It can assess whether the information from adjacent stations contributes to improving the power forecasting of the target station, which is obviously an effective manner to select the most valuable adjacent stations.

The expression of transfer entropy is formulated by

$$T_{M\to N} = H\left(N_t \mid N_{t-1:t-L}\right) - H\left(N_t \mid N_{t-1:t-L}, M_{t-1:t-L}\right) \quad (1)$$

where $T_{M\to N}$ represents the improvement of the forecasting of sequence N due to the historical information from the sequence M. $H(N_t \mid N_{t-1:t-L})$ denotes the uncertainty in forecasting N_t when only the historical information of the sequence N is considered, while $H(N_t \mid N_{t-1:t-L}, M_{t-1:t-L})$ represents the uncertainty in forecasting N_t when both the historical information of sequences N and M are incorporated. If $T_{M\to N}$ is greater than zero, it indicates that the historical information of sequence M is beneficial to improving the forecasting of N_t.

Transfer entropy allows the model to select adjacent stations according to forecasting relevance, rather than relying on crude geographic proximity. High transfer entropy value indicates that data from this adjacent station is beneficial to improving the prediction performance of the target station. It is an effective manner to facilitate the forecasting performance.

2.2 Graph-structured modeling

Graph-structured modeling provides an effective way to represent complex relational systems, where nodes denote entities with intrinsic attributes and edges encode the interactions or dependencies among them[21]. By explicitly modeling these relationships, graph-based representations enable the learning framework to capture both local and global structural information inherent in the data.

In neural graph-based models, node representations are often updated via message-passing mechanisms,

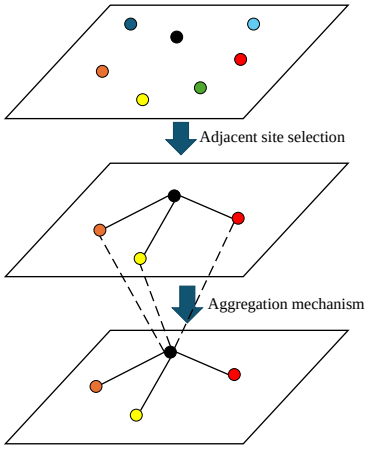

Fig. 1 The processing procedure of graph data.

which can be formulated as:

$$h_v^{(l)} = \sigma\left(H^{(l)} * \sum_{u \in \mathcal{N}(v)} h_u^{(l-1)}\right), \quad (2)$$

where $h_v^{(l)}$ denotes the representation of node v at the l-th layer, $\sigma(\cdot)$ is a nonlinear activation function, $H^{(l)}$ is a trainable transformation matrix, and $\mathcal{N}(v)$ represents the neighborhood of node v. Through such hierarchical aggregation, the model captures structural dependencies and interaction patterns within the graph.

In this work, each node corresponds to a PV station, as illustrated in Fig. 1. Transfer entropy analysis is employed to identify influential neighboring stations, based on which a graph structure is constructed. The resulting edges explicitly encode the dependency between the target station and its adjacent stations, enabling the proposed model to effectively exploit inter-station relationships for PV power forecasting.

2.3 Transformer

The Transformer model is built on an encoder-decoder architecture, which has illustrated remarkable efficacy in time series forecasting[22]. It employs self-attention and cross-attention mechanisms to model the relationships between the source and target sequences. Self-attention, as the core component of the Transformer, enables the model to capture long-range dependencies within the sequence.

The attention coefficient is calculated by

$$Attn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where

$$Q = W_q P \quad (4)$$

$$K = W_k P \quad (5)$$

$$V = W_v P \quad (6)$$

in which W_q, W_k, W_v are trainable weight matrices, P is the input, d_k is the dimension of model.

The relationships between time series are generally complex, and a single attention head may not be sufficient to capture all such relationships. To address this issue, multi-head attention was introduced to capture different relationships simultaneously[22]. The Transformer architecture effectively combines both global and local attention mechanisms, allowing for parallel computation and improving model efficiency.

2.4 Problem statement

Given n PV stations, the input data $X = [x_1, x_2, \ldots, x_t, \ldots, x_n] \in \mathbb{R}^{n \times s \times f}$ represents the historical data of all the stations. And $x \in \mathbb{R}^{s \times f}$ corresponds to the data of a single PV station and x_t represents the target station. f is the number of features and s is the time steps for each station.

This paper aims to establish an accurate PV forecasting model that predicts the target station's output over a future horizon of length p. The output is denoted by $x'_p \in \mathbb{R}^{p \times f}$, which represents the forecasting result for the target station. To enhance the forecasting performance, the adjacent stations are first chosen to construct valuable data. Then, a model based on Transformer and graph modeling is investigated to characterize spatio-temporal relationships between PV stations. The task can be briefly summarized as follows.

$$X \xrightarrow{Adjacent\ station\ selection} X' \xrightarrow{Model\ prediction} x'_p \quad (7)$$

where X' represents the selected data.

3 Methodology

In this section, the comprehensive architecture of the proposed ns_Transraph method is summarized. Subsequently, the significant modules of ns_Transraph are described in detail.

3.1 Overall architecture and ns_Transraph

There are three main steps in the PVPF process. First, the data is preprocessed to enhance the quality, including removing outliers, interpolating data, and normalizing data. Then the adjacent PV stations are selected through transfer entropy, which would provide beneficial information for the target PV station. Eventually, the data from the target PV station and the selected stations are utilized to establish the proposed ns_Transraph model, as summarized in Fig. 2. For the input, x_t represents the data of the target PV station, x_o represents the data of adjacent PV stations and x_p represents the real data which should be forecasted.

The proposed ns_Transraph adopts an encoder-decoder architecture. It integrates the Transformer with graph structure and fully utilizes information from adjacent stations. It characterizes the complex spatio-temporal relationships between the time series of adjacent PV stations.

(1) To integrate the spatio-temporal information of adjacent PV stations, in the encoder part, a novel implicit graph based on cross attention aggregation mechanism is designed to extract the spatio-temporal relationships of the adjacent PV stations.

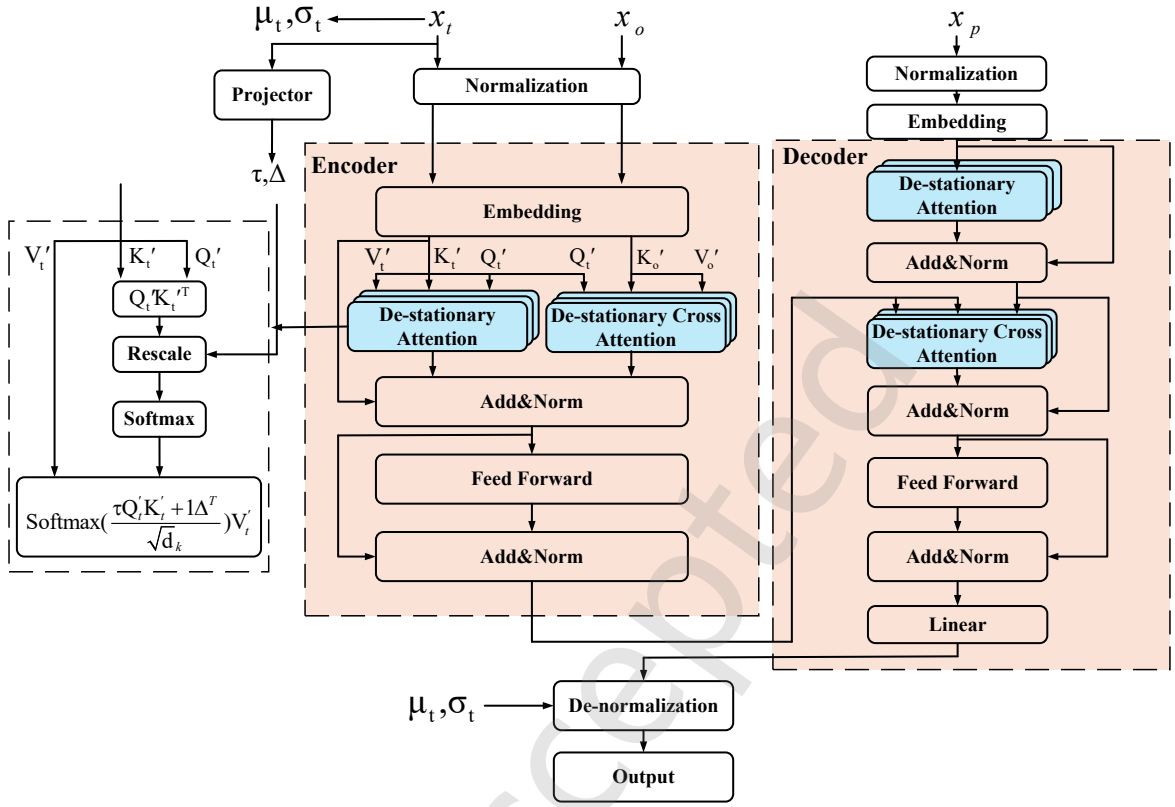

Fig. 2 The architecture of ns_Transraph.

(2) To focus on the nonstationarity of the target PV station, a graph-inspired de-stationary cross attention is introduced in the encoder and decoder part to aggregate the information of adjacent power stations while recovering the non-stationarity of the original target PV station time series.

(3) To take advantage of the encoder-decoder structure, in the other part of the model, the model follows the traditional architecture of the Transformer. The input of the decoder is the forecasting target and the historical data of the target PV station.

3.2 Proposed implicit graph modeling spatial cross attention aggregation mechanism

Many models fail to preserve the data structure because flattening and concatenating the data can result in the loss of temporal dependencies. To integrate the information from the adjacent PV stations, a cross-attention aggregation mechanism is proposed to capture the spatio-temporal relationships between the target station and its adjacent stations. Unlike the explicit GNN, our work is based on implicit graph modeling embedded in the encoder part of Transformer.

As depicted in Fig. 1, each adjacent PV station represents a node in the graph-structured data. The feature of each node is modeled as a two-dimensional tensor, encompassing multiple features and time stamps. This approach eliminates the need for flattening and concatenating tensors during training. As a result, it preserves the temporal correlations between sequences. Furthermore, by treating the nodes as PV stations rather than discrete time stamps, the model leverages the full potential of the graph to capture spatial dependencies between the nodes. This approach mitigates issues related to accuracy degradation and the excessive parameterization that can arise from an overly large number of graph layers.

The cross-attention coefficient is calculated through the target station and its adjacent stations, as depicted in Fig. 3. The self-attention of the target station is calculated by

$$A_t = Softmax\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t \tag{8}$$

where A_t represents the self-attention coefficient of the target station, Q_t, K_t, V_t can be calculated as Eqs.(4)-

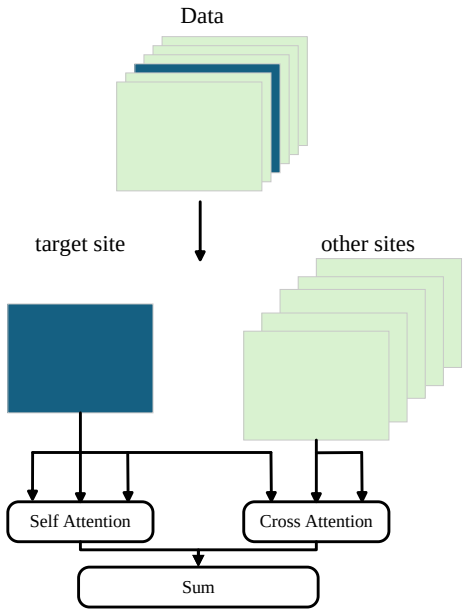

Fig. 3 Cross attention aggregation mechanism.

(6) respectively.

The cross-attention between the target station and its adjacent stations is computed by

$$A_{ti} = Softmax\left(\frac{Q_t K_i^T}{\sqrt{d_k}}\right) V_i \tag{9}$$

where A_{ti} represents the cross-attention coefficient between the target station t and its adjacent station i, K_i, V_i can be calculated as Eqs. (5)–(6). Then, the result of aggregation is expressed by

$$A = Softmax\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t + \sum_{i=1, i\neq t}^{n} Softmax\left(\frac{Q_t K_i^T}{\sqrt{d_k}}\right) V_i \tag{10}$$

Through Eq (10), the information of adjacent PV stations is successfully aggregated. Although simple addition is employed for aggregation here, the summation does not represent a simple linear aggregation of neighboring stations. In Eq (10), cross-attention is employed instead of the traditional explicit graph neural network aggregation mechanism. Using an implicit graph modeling approach, the information of neighboring power stations is weighted in the form of attention, and a nonlinear relationship is extracted through softmax, successfully modeling complex spatiotemporal relationships.

3.3 De-stationary enhancement for cross-Attention aggregation

Due to the significant non-stationary characteristics, stationarization is commonly employed to enhance forecasting performance. However, Liu et al.[23] showed that stationarization may suppress important non-stationary patterns in the original time series. To address this issue, a de-stationary attention mechanism was proposed to prevent over-stationarization[23], thereby improving the forecasting accuracy. In addition, the stacking of multiple graph layers can lead to excessive smoothing, resulting in the loss of node discrimination and the dilution of information pertinent to the target node.

To address these challenges, a de-stationary attention mechanism is introduced, and a graph-inspired de-stationary cross-attention module is proposed to preserve the non-stationary characteristics of the target node during attention-based spatial aggregation. Different from prior studies that applied de-stationary modeling exclusively to self-attention, the proposed method incorporates de-stationary factors directly into the cross-attention mechanism, explicitly regulating the aggregation of spatial information from neighboring stations. This design is motivated by the observation that graph aggregation tends to introduce over-stationarization effects, which the proposed module effectively mitigates.

The original time series of target station X is normalized as $X' = (X - \mathbf{1}\mu_x)/\sigma_x$, where $\mathbf{1} \in \mathbb{R}^{s\times 1}$ is an all-ones vector. μ_Q can be calculated by

$$\mu_Q = W_Q \mu_x \tag{11}$$

The softmax operation of the original sequence can be expressed as follows.

$$Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) = Softmax\left(\frac{\sigma_x^2 Q' K'^T + \mathbf{1}\left(\mu_Q^T K^T\right)}{\sqrt{d_k}}\right) \tag{12}$$

where $\tau = \sigma_x^2$ and $\triangle = K\mu_Q$, two multilayer perceptron layers are constructed to learn them from the original time series. Eventually, they are recovered in the attention layer as follows.

$$Attn(Q', K', V', \tau, \triangle) = Softmax\left(\frac{\tau Q' K'^T + \mathbf{1}\triangle^T}{\sqrt{d_k}}\right) V' \tag{13}$$

To preserve the non-stationary characteristics of the target station during spatial aggregation, a de-stationary

cross-attention mechanism is incorporated into the graph-modeling and decoder part. The de-stationary factors are explicitly introduced into the implicit graph modeling module and decoder's cross-attention, aiming to re-calibrate the aggregated representations and preserve the non-stationary dynamics of the target station. This design enables the model to benefit from spatial information while maintaining sensitivity to abrupt variations of target station.

4 Experiments

Experiments are conducted based on real-world time series to evaluate the performance of the proposed ns_Transraph method.

4.1 Datasets and experiment set

4.1.1 Data description

Two real-world photovoltaic power generation datasets are utilized to verify the superiority and robustness of the proposed model.

•Dataset 1: It contains 9 distributed PV stations with longitude and latitude information from Fujian Province, China. The geographic information of 9 stations is shown in Table 1. The target station is Station 1. Then the distance between the target station and each station is calculated and listed in Table 1. For each station, the data covers a four-month period from January 3, 2022 to April 30, 2023, with a temporal resolution of 15 minutes, yielding a total of 46,368 records after preprocessing. The data is divided into training set, verification set, and testing set in the ratio of 7:1:2. Each record contains seven features per station, including temperature, pressure, rainfall, irradiance, wind speed, cloud cover, and power. The dataset used in this study is publicly available at https://github.com/classKdsnnModule/History-PV-power-data-of-distributed-PV-stations-in-Fujian-China-with-loaction.git.

Table 1 Geographic information of stations in Dataset 1

Station	Longitude	Latitude	Distance (km)
1	119.2186	26.04293	0
2	115.1245	24.69532	185.38
3	117.0021	25.1125	244.70
4	117.8549	26.74467	157.20
5	120.0223	26.87252	122.30
6	119.156	25.44923	65.60
7	118.8613	25.13104	108.90
8	117.5771	26.28068	165.50
9	117.7405	24.07764	263.80

• Dataset 2: It contains 12 distributed PV stations without longitude and latitude information. For each station, the dataset covers a four-month period from June 1 to September 30, 2023, with a temporal resolution of 15 minutes, yielding a total of 11,222 records after preprocessing. The data is divided into training set, verification set, and testing set in the ratio of 7:1:2. Each record contains seven features per station, including temperature, pressure, rainfall, irradiance, wind speed, cloud cover and power. The target station is Station 1.

4.1.2 Experiment set

The forecasting performance of the proposed method is evaluated through four popular metrics, including mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2). $x_p^{(i)}$ and $x_p'^{(i)}$ represent the true and forecast value of the i-th time step respectively, $\bar{x}_p$ is the average of x_p, and p represents the total number of forecast time steps. The metrics are expressed as follows.

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{p}\sum_{i=1}^{p}\left|x_p^{(i)} - x_p'^{(i)}\right| \\ \mathrm{MSE} &= \frac{1}{p}\sum_{i=1}^{p}\left(x_p^{(i)} - x_p'^{(i)}\right)^2 \\ \mathrm{MAPE} &= \frac{100\%}{p}\sum_{i=1}^{p}\left|\frac{x_p^{(i)} - x_p'^{(i)}}{x_p^{(i)}}\right| \\ \mathrm{R}^2 &= 1 - \frac{\sum_{i=1}^{p}\left(x_p^{(i)} - x_p'^{(i)}\right)^2}{\sum_{i=1}^{p}\left(x_p^{(i)} - \bar{x}_p\right)^2} \end{aligned} \tag{14}$$

The experiment was conducted on a personal computer (CPU: AMD Ryzen 5 7600X 5.3GHz, GPU: NVIDIA GeForce RTX 4060 8GB, RAM: 32GB). The code was written based on Pytorch 2.3.0 for Python 3.12. To fully display the advantages of the proposed ns_Transraph method, Transformer [22], Informer[24], Autoformer[25], LSTM[26], RNN[27], STGCN[28] and GraphWaveNet[29], are chosen as comparison models, utilizing the data of the past 24 hours to forecast with different time horizons.

The parameters of models are summarized in Table 2. For Transformer and its variants, all parameters are set the same. d_model is the dimension of the model, n_head is the number of attention heads, e_layer is the number of encoder layer, d_layer is the number of decoder layer, d_ff is the dimension of feed forward layer. For LSTM and RNN, the size of hidden_size is

Table 2 Parameters of models

Model	Parameters
Transformer	d_model = 512, n_head = 8, e_layer = 2, d_layer = 1, d_ff = 2048
Informer	d_model = 512, n_head = 8, e_layer = 2, d_layer = 1, d_ff = 2048
Autoformer	d_model = 512, n_head = 8, e_layer = 2, d_layer = 1, d_ff = 2048
LSTM	hidden_size = 512, num_layer = 2
RNN	hidden_size =512, num_layer = 2
STGCN	num_block = 2, spatial_channel = 16
GraphWaveNet	n_block = 4,n_layer = 2,dilation_channel = 32, residual_channel = 32, skip_channel = 256 end_channel = 512
ns_Transraph	d_model = 512, n_head = 8, e_layer = 2, d_layer = 1, d_ff = 2048

the same as that of d_model, and num_layer is equal to e_layer. For STGCN, num_block is the number of STGCN blocks, spatial_channel is the number of spatial patterns extracted by GCN. For GraphWaveNet, n_block is the number of WaveNet blocks, n_layer is the number of layers in each block, dilation_channel is the number of channels in the convolutional layer of the temporal convolutional module, residual_channel is the feature dimension maintained by the residual connections of each layer, skip_channel is the number of jumper output channels for each layer, end_channel is the number of channels of the convolutional layer used in the final output module.

4.2 Experiments on Dataset 1

4.2.1 Selection of adjacent stations

Transfer entropy analysis between the target station and other stations is conducted, and the results are recorded in Table 3. Considering the computational complexity and the potential impact of low information gain stations on noise, the percentage threshold of the relative maximum TE value is used to select the adjacent power stations.

$$\tau = \frac{TE_i}{TE_{max}} \tag{15}$$

where TE_i represents the transfer entropy of Station i, and TE_{max} represents the maximum transfer entropy of other stations. Those sites whose τ greater than 60% will be selected as adjacent sites. For dataset 1, Station 5 and Station 6 are selected as adjacent stations.

According to Table 3 and distance information in Table 1, it can be observed that Station 7 is geographically closer to Station 1, but Station 5 located to the northeast, exhibits higher transfer entropy. This result is attributable to the directional precedence of the transmission of information from Station 5 to Station 1.

Table 3 Transfer entropy result of Dataset 1

Station	2	3	4	5	6	7
Transfer entropy	0.0181	0.0131	0.0161	0.0260	0.0342	0.0173
Station	8	9				
Transfer entropy	0.0148	0.0131				

4.2.2 Main result analysis

The evaluation metrics of different DL methods on Dataset 1 are listed in Table 4. Table 4 shows that MSE consistently increases as the forecasting horizon extends from 15 minutes to 4 hours for all models, reflecting the growing difficulty of long-term prediction. In short-term forecasting, ns_Transraph achieves an MSE of 0.047 at the 15-minute horizon, which is comparable to the lowest MSE of 0.046 reported by GraphWaveNet, indicating that the proposed non-stationary modeling does not sacrifice short-term accuracy. At 30-minute horizon, ns_Transraph further reduces the MSE to 0.058, outperforming all baseline methods.

As the prediction horizon increases, the advantage of ns_Transraph in terms of MSE becomes increasingly pronounced. At the 1-hour horizon, ns_Transraph achieves the lowest MSE of 0.078, improving over the second-best result of 0.079. This performance gap further widens at longer horizons. At 4-hour horizon, it reduces the MSE to 0.148 compared with 0.169 achieved by the second-best model. Although some baseline methods obtain slightly lower MAE or MAPE values at specific horizons, ns_Transraph consistently demonstrates slower MSE growth and superior long-term stability, highlighting its effectiveness in mitigating error accumulation and modeling non-stationary temporal dynamics.

4.2.3 Verification and sensitivity analysis of adjacent station selection

This section evaluates the robustness of the proposed adjacent station selection strategy and clarifies its rationale. Stations are selected based on their transfer entropy with the target station, rather than purely geographic proximity. And the experiment about the sensitivity analysis of transfer entropy threshold is conducted.

Based on the transfer entropy analysis in Table 3 and geographic information in Table 1, Stations 2, 5, 6,

Table 4 Experiment results of Dataset 1

Model	Metrics	Horizon 15min	Horizon 30min	Horizon 1h	Horizon 2h	Horizon 4h
Transformer	MSE	0.047	0.061	0.084	0.104	0.169
	MAE	0.094	0.113	0.153	0.169	**0.195**
	MAPE	0.683	0.803	1.104	1.304	1.505
	R^2	0.940	0.924	0.895	0.870	0.789
Informer	MSE	0.053	0.059	0.079	0.109	0.176
	MAE	0.109	0.112	0.131	**0.153**	0.210
	MAPE	0.909	0.949	**1.038**	**1.117**	1.322
	R^2	0.933	0.926	0.901	0.864	0.780
Autoformer	MSE	0.130	0.146	0.187	0.274	0.369
	MAE	0.259	0.274	0.302	0.359	0.497
	MAPE	1.156	1.530	1.517	1.500	1.916
	R^2	0.837	0.818	0.766	0.658	0.538
LSTM	MSE	0.046	0.062	0.087	0.137	0.219
	MAE	0.092	0.113	0.141	0.181	0.228
	MAPE	0.752	1.041	1.258	1.366	1.531
	R^2	0.942	0.923	0.891	0.829	0.726
RNN	MSE	0.051	0.066	0.089	0.138	0.225
	MAE	0.111	0.128	0.153	0.190	0.241
	MAPE	0.792	1.106	1.203	1.201	1.476
	R^2	0.936	0.917	0.889	0.827	0.718
STGCN	MSE	0.101	0.103	0.116	0.126	0.175
	MAE	0.186	0.173	0.195	0.199	0.246
	MAPE	1.212	1.362	1.340	1.410	1.661
	R^2	0.874	0.871	0.855	0.842	0.780
GraphWaveNet	MSE	**0.046**	0.060	0.082	0.120	0.189
	MAE	**0.091**	**0.107**	**0.129**	0.162	0.205
	MAPE	**0.691**	0.920	1.046	1.237	**1.259**
	R^2	**0.942**	0.925	0.898	0.850	0.763
ns_Transraph	MSE	0.047	**0.058**	**0.078**	**0.103**	**0.148**
	MAE	0.101	0.114	0.150	0.164	0.201
	MAPE	0.705	**0.843**	1.140	1.210	1.384
	R^2	0.940	**0.928**	**0.902**	**0.871**	**0.815**

and 7 are considered candidate neighboring stations for Station 1. Different combinations of these stations are used to evaluate the impact of spatial information on forecasting performance.

Table 5 Verification of different adjacent stations

Adjacent stations	MSE	MAE	MAPE	R^2
7	0.162	0.218	1.578	0.797
5	0.154	0.199	1.458	0.808
5, 7	0.172	0.206	1.613	0.785
5, 6	0.148	0.201	1.384	0.815
6, 7	0.164	0.201	1.587	0.794

As shown in Table 5, the MSE of the model with Station 7 as the adjacent station is 0.162, and the MSE of the model with Station 5 as the adjacent Station is 0.154. This is contrary to the relationship with geographical distance, but it is consistent with the results of the transfer entropy analysis. This indicates that transfer entropy can more quantitatively represent the information transmission between adjacent stations compared to simple geographical distance. When two stations are considered, the model with stations 5 and 6 as adjacent stations is even better than the model with stations 6 and 7 as adjacent stations.

Table 6 Sensitivity analysis of transfer threshold

TE Threshold (τ)	Adjacent stations	MSE	R^2
0.0222 (65%)	5, 6	0.148	0.815
0.0205 (60%)	5, 6	0.148	0.815
0.0188 (55%)	5, 6	0.148	0.815
0.0171 (50%)	5, 6, 2, 7	0.157	0.804

The sensitivity of transfer threshold is analyzed and the results are listed in Table 6. The transmission entropy threshold changes with τ. As τ decreases, fewer stations would be filtered out. If τ is lower than 55%, the number of adjacent power stations increases from two to four. Under this combination, the performance of the model declines. Since an excessive number of adjacent power stations can introduce noise and cause information duplication, the performance of the model may be degraded. This indicates that setting the maximum τ at 60% as the threshold is reasonable and can effectively select adjacent sites.

In conclusion, the aforementioned analysis illustrates that the proposed transfer-entropy-based adjacent station selection strategy is rational and robust. It effectively captures the most informative neighboring stations and outperforms purely geographically-based selection. Moreover, the performance is relatively robust to the exact number of stations or TE threshold within a reasonable range.

4.3 Experiments on Dataset 2

4.3.1 Selection of adjacent stations

Since the geographic information of these stations is not available, it is difficult to filter adjacent stations through the traditional geographical distance. To select the station that provides the greatest information gain to the target station, transfer entropy is used to quantify the gain of the remaining station to the target station and the results are recorded in Table 7. For dataset 2, Station 4 and Station 10 are selected as adjacent stations according to Eq (15).

Table 7 Transfer entropy result of Dataset 2

Station	2	3	4	5	6	7
Transfer entropy	0.0169	0.0136	0.0205	0.0178	0.0124	0.0066
Station	8	9	10	11	12	
Transfer entropy	0.0150	0.0085	0.0316	0.0184	0.0168	

4.3.2 Main result analysis

The evaluation metrics of different DL methods on

Table 8 Experiment results of Dataset 2

Model	Metrics	Horizon 15min	Horizon 30min	Horizon 1h	Horizon 2h	Horizon 4h
Transformer	MSE	0.026	0.063	0.104	0.197	0.232
	MAE	0.111	0.159	0.218	0.293	0.325
	MAPE	0.280	0.431	0.566	**0.610**	0.849
	R^2	0.975	0.940	0.901	0.814	0.779
Informer	MSE	0.045	0.045	0.122	0.142	0.187
	MAE	0.144	0.143	0.226	0.236	0.310
	MAPE	0.558	0.458	0.673	0.933	1.067
	R^2	0.957	0.958	0.884	0.865	0.822
Autoformer	MSE	0.156	0.322	0.289	0.296	0.637
	MAE	0.320	0.428	0.413	0.417	0.610
	MAPE	0.924	1.584	1.477	1.351	1.635
	R^2	0.852	0.695	0.726	0.719	0.394
LSTM	MSE	0.025	0.046	0.085	0.152	0.337
	MAE	0.107	0.144	0.196	0.245	0.360
	MAPE	0.440	0.625	0.731	0.893	0.911
	R^2	0.976	0.956	0.920	0.856	0.680
RNN	MSE	**0.016**	**0.033**	0.072	0.174	0.292
	MAE	**0.069**	**0.117**	**0.175**	0.268	0.342
	MAPE	**0.241**	0.423	0.730	0.894	1.034
	R^2	**0.985**	**0.968**	0.932	0.835	0.723
STGCN	MSE	0.030	0.033	**0.063**	0.208	0.331
	MAE	0.125	0.133	0.231	0.327	0.373
	MAPE	0.312	**0.421**	0.680	1.116	1.225
	R^2	0.971	0.950	0.929	0.882	0.808
GraphWaveNet	MSE	0.023	0.035	0.094	0.113	0.190
	MAE	0.112	0.130	0.211	0.208	0.273
	MAPE	0.398	0.495	0.646	0.712	1.134
	R^2	0.978	0.967	0.911	0.893	0.819
ns_Transraph	MSE	0.018	0.039	0.070	**0.091**	**0.142**
	MAE	0.083	0.136	0.177	**0.206**	**0.246**
	MAPE	0.267	0.452	**0.558**	0.711	**0.819**
	R^2	0.983	0.963	**0.933**	**0.914**	**0.865**

Dataset 2 are shown in Table 8. Bold denotes the best results and underline denotes the second-best results.

It can be observed that ns_Transraph consistently achieves superior or highly competitive performance across different prediction horizons, especially for the medium and long-term horizons such as 2 hours and 4 hours. At the forecasting horizons of 2 and 4 hours, ns_Transraph achieves MSE of 0.091 and 0.142, representing improvements of 24.2% and 24.1% respectively over the second-best model. This illustrates that the proposed model successfully extracts the spatio-temporal information from adjacent PV stations. As a result, it significantly improves the prediction accuracy of the target station, particularly in medium and long-term forecasting.

For short-term horizons, the performance of ns_Transraph is slightly inferior to that of the baseline model, but the gap is not significant. For instance, at the 15-minute forecasting horizon, it achieves an MSE of 0.018, slightly higher than that of RNN, ranking as the second-best model. At the 1-hour forecasting horizon, ns_Transraph records an MSE of 0.070, surpassing GraphWaveNet with an MSE of 0.094 and approaching STGCN with an MSE of 0.063, while achieving lower MAE and MAPE.

The performance of all models degrades as the prediction horizon increases. For LSTM and RNN, RNN outperforms LSTM in the short-term horizon of prediction. As the horizon of prediction increases, the advantages of LSTM in long sequence prediction are gradually being illustrated. However, LSTM and RNN degrade more sharply than the proposed model due to the circular structure. Similarly, Transformer and its variants perform better than LSTM, but degrade faster than the proposed ns_Transraph because only the history information of the target station is considered. The performance of Autoformer degrades sharply because the data information is limited and is not able to capture seasonal periodicity. For GraphWaveNet and STGCN, in short-term modeling, the convolutional graph structure combined with the information of adjacent stations performs well. But the historical information is gradually diluted in long-term modeling.

4.3.3 Ablation experiment

To illustrate the effectiveness of the proposed modules, comprehensive ablation experiments are conducted on Dataset 2. The experiments compare the performance of the vanilla Transformer with Transraph without the de-stationary module across multiple time horizons. Additionally, the results of ns_Transformer are included to highlight the advantages of the cross-attention aggregation mechanism. Notably, the input data for all models are sourced from three adjacent PV stations, ensuring a consistent baseline to evaluate the models' spatio-temporal extraction capabilities. For the Transformer and ns_Transformer, the encoder and decoder input sizes are adjusted to allow the models to autonomously extract spatio-temporal relationships from the raw data. The experimental results are presented in Table 9.

The results clearly indicate that Transraph consistently outperforms the vanilla Transformer across nearly all time horizons. For instance, at a 15-minute horizon, Transraph achieves an MSE of 0.023 compared to the Transformer's 0.046, and an R^2 score of 0.978 and 0.957, illustrating superior forecast accuracy. Similarly, ns_Transraph exhibits significant improvements over ns_Transformer, particularly in longer horizons. At the 4-hour horizon, ns_Transraph attains an MSE of 0.142 and an R^2 of 0.865, while ns_Transformer records an MSE of 0.156 and an R^2

Table 9 Ablation experiment

Model	Metrics	Horizon 15min	Horizon 30min	Horizon 1h	Horizon 2h	Horizon 4h
Transformer	MSE	0.046	0.089	0.118	0.157	0.251
	MAE	0.151	0.184	0.220	0.268	0.313
	MAPE	0.347	0.394	0.540	0.632	0.824
	R^2	0.957	0.916	0.887	0.851	0.761
ns_Transformer	MSE	0.021	0.039	0.067	0.109	0.156
	MAE	0.089	0.131	0.172	0.211	0.250
	MAPE	0.255	0.351	0.535	0.692	0.965
	R^2	0.980	0.963	0.935	0.896	0.852
Transraph	MSE	0.023	0.062	0.090	0.186	0.214
	MAE	0.103	0.161	0.223	0.308	0.331
	MAPE	0.284	0.469	0.629	0.611	1.000
	R^2	0.978	0.941	0.914	0.824	0.797
ns_Transraph	MSE	0.018	0.039	0.070	0.091	0.142
	MAE	0.083	0.136	0.177	0.206	0.246
	MAPE	0.267	0.452	0.558	0.711	0.819
	R^2	0.983	0.963	0.933	0.914	0.865

of 0.852. These results show that simply expanding model inputs is insufficient. Specialized mechanisms are needed to extract spatio-temporal relationships.

Furthermore, the superior performance of ns_Transraph over Transraph highlights the efficacy of the cross-de-stationary attention mechanism in capturing complex spatio-temporal dependencies. For example, at the 30-minute horizon, ns_Transraph achieves an MAE of 0.136 and a MAPE of 0.452, compared to Transraph's MAE of 0.161 and MAPE of 0.469. This improvement validates the hypothesis that integrating de-stationary modules and cross-attention mechanisms enhances the ability of the model to handle non-stationarity and spatially correlated data.

In summary, the ablation experiments confirm that the proposed Transraph architecture, particularly with the inclusion of the cross-de-stationary attention mechanism, significantly enhances the extraction of spatio-temporal relationships between adjacent PV stations. The outperformance of Transraph and ns_Transraph over corresponding models highlights the necessity of the modules proposed for PVPF.

4.4 Computational complexity analysis

The training time and testing time of all models are summarized in Table 10. It can be observed that the proposed ns_Transraph exhibits a moderate training cost. Its training time of two datasets are 117.13s and 85.93s, which is higher than that of lightweight recurrent models such as RNN, LSTM and STGCN. However, it is obviously lower than more computationally intensive architectures, including Informer, Autoformer and GraphWaveNet. Overall, ns_Transraph maintains a reasonable balance between model complexity and training efficiency among Transformer-based and graph-based approaches.

Table 10 Computational complexity

Model	Dataset	Training time(s)	Testing time(s)
Transformer	1	291.80	61.71
	2	78.18	17.89
Informer	1	355.98	87.65
	2	118.04	20.96
Autoformer	1	625.48	89.03
	2	168.40	21.7
LSTM	1	268.26	63.43
	2	53.10	12.72
RNN	1	178.38	63.95
	2	47.47	13.89
STGCN	1	129.28	49.89
	2	44.60	11.71
GraphWaveNet	1	1076.44	100.21
	2	347.51	25.04
ns_Transraph	1	394.82	75.75
	2	85.93	18.62

In terms of inference efficiency, ns_Transraph provides competitive testing performance. Its total testing time of two datasets are 75.75s and 18.62s, which corresponds to an average inference latency of approximately 8.18ms to 9.3ms per prediction. This latency is slightly higher than that of the vanilla Transformer with 6.67ms to 8.9 ms per prediction, but is clearly lower than GraphWaveNet with 10.82 ms to 12.7ms per prediction. Besides, it is comparable to Informer and Autoformer, whose inference latency is around 10.8ms to 10.9ms per prediction. Under the same hardware configuration, the inference latency gap between ns_Transraph and other attention-based models remains limited, indicating that the proposed method can satisfy the real-time requirements of PVPF with a 5-minute sampling interval.

4.5 Discussion and limitation

Although ns_Transraph alleviates over-stationarity via the de-stationary module, the aggregation of spatiotemporal features may still suffer from spectral bias, where high-frequency signals are suppressed.

Inspiringly, recent dense prediction methodologies have tackled similar feature degradation issues through the lens of frequency analysis. Studies on frequency aliasing [30, 31] highlight the necessity of preserving high-frequency details during sampling to maintain feature fidelity. Moreover, dynamic frequency modulation techniques [32, 33, 34] provide adaptive mechanisms to balance spectral components,

effectively countering the inherent low-pass filtering tendency of deep networks. Additionally, frequency-aware fusion[35] offers a robust strategy for maintaining feature consistency in complex boundaries.

Drawing on these advanced paradigms, future work will explore explicit frequency-adaptive modules. Integrating frequency-domain constraints may further improve the model's robustness against volatility and eliminate the residual over-smoothing effects in long-term forecasting.

5 Conclusion

In this paper, ns_Transraph, a novel framework combining Transformer and implicit graph modeling, is proposed to address the spatio-temporal relationship in ultra-short-term PVPF. The model includes two key innovations, namely a cross-attention aggregation mechanism that simultaneously captures spatial and temporal relationships, and a de-stationary module that preserves the nonstationarity of target PV stations while mitigating over-smoothing in graph layers. The experiments illustrate that ns_Transraph outperforms other models in PVPF, especially in long-term horizons. The optimal selection of adjacent stations revealed a trade-off between information gain and noise. It also reveals the importance of geographical coverage when adjacent stations are selected. Moreover, the de-stationary module enhanced the robustness of the model under complex weather conditions, making it suitable for grid dispatching.

Future work can attempt to address the issue of non-stationarity in data from a frequency domain perspective.

Acknowledgment

This work was supported by State Grid Corporation Headquarters Technology Project [grant number 4000-202316456A-3-2-ZN].

References

[1] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif, A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization, *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109792, 2020.

[2] M. J. Mayer and D. Yang, Pairing ensemble numerical weather prediction with ensemble physical model chain for probabilistic photovoltaic power forecasting, *Renewable and Sustainable Energy Reviews*, vol. 175, p. 113171, 2023.

[3] G. Xu, H. Wang, S. Ji, Y. Ma, and Y. Feng, MPformer: A transformer-based model for earthen ruins climate prediction, *Tsinghua Science and Technology*, vol. 29, no. 6, pp. 1829–1838, 2024.

[4] Z. Jiang, Z. Ning, H. Miao, and L. Wang, STDNet: A spatio-temporal decomposition neural network for multivariate time series forecasting, *Tsinghua Science and Technology*, vol. 29, no. 4, pp. 1232–1247, 2024.

[5] O. Akbarzadeh, S. Hamzehei, H. Attar, A. Amer, N. Fasihihour, M. R. Khosravi, and A. A. Solyman, Heating-cooling monitoring and power consumption forecasting using LSTM for energy-efficient smart management of buildings: A computational intelligence solution for smart homes, *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 143–157, 2024.

[6] S. Srivastava and S. Lessmann, A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data, *Solar Energy*, vol. 162, pp. 232–247, 2018.

[7] K. Wang, X. Qi, and H. Liu, Photovoltaic power forecasting based LSTM-convolutional network, *Energy*, vol. 189, p. 116225, 2019.

[8] L. Yin, X. Cao, and D. Liu, Weighted fully-connected regression networks for one-day-ahead hourly photovoltaic power forecasting, *Applied Energy*, vol. 332, p. 120527, 2023.

[9] Q. liu, Y. li, H. jiang, Y. chen, and J. Zhang, Short-term photovoltaic power forecasting based on multiple mode decomposition and parallel bidirectional long short term combined with convolutional neural networks, *Energy*, vol. 286, p. 129580, 2024.

[10] J. López Lorente, X. Liu, and D. J. Morrow, Spatial aggregation of small-scale photovoltaic generation using voronoi decomposition, *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2677–2686, 2020.

[11] Y. Wen, S. Pan, X. Li, Z. Li, and W. Wen, Improving multi-site photovoltaic forecasting with relevance amplification: Deepfedformer-based approach, *Energy*, vol. 299, p. 131479, 2024.

[12] W. Lai, Z. Zhen, F. Wang, W. Fu, J. Wang, X. Zhang, and H. Ren, Sub-region division based short-term regional distributed PV power forecasting method considering spatio-temporal correlations, *Energy*, vol. 288, p. 129716, 2024.

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

[14] K. Song, M. Kim, and H. Kim, Graph-based large scale probabilistic PV power forecasting insensitive to space-time missing data, *IEEE Transactions on Sustainable Energy*, vol. 16, no. 1, pp. 160–173, 2025.

[15] M. Bai, Z. Zhou, J. Li, Y. Chen, J. Liu, X. Zhao, and D. Yu, Deep graph gated recurrent unit network-based spatial–temporal multi-task learning for intelligent information fusion of multiple sites with application in short-term spatial–temporal probabilistic forecast of

photovoltaic power, *Expert Systems with Applications*, vol. 240, p. 122072, 2024.

[16] X. Zhang, R. Gao, C. Zhu, C. Liu, and S. Mei, Ultra-short-term prediction of regional photovoltaic power based on dynamic graph convolutional neural network, *Electric Power Systems Research*, vol. 226, p. 109965, 2024.

[17] Y. Yang, Y. Liu, Y. Zhang, S. Shu, and J. Zheng, DEST-GNN: A double-explored spatio-temporal graph neural network for multi-site intra-hour PV power forecasting, *Applied Energy*, vol. 378, p. 124744, 2025.

[18] Y. Gao, L. Liang, T. Su, and M. Pan, An embedded spatiotemporal hybrid model integrating multi-graphs and attention-driven fusion for single- and multi-site photovoltaic power forecasting, *Energy Conversion and Management*, vol. 336, p. 119897, 2025.

[19] M. Bai, G. Zhou, P. Yao, F. Dong, Y. Chen, Z. Zhou, X. Yang, J. Liu, and D. Yu, Deep multi-attribute spatial–temporal graph convolutional recurrent neural network-based multivariable spatial–temporal information fusion for short-term probabilistic forecast of multi-site photovoltaic power, *Expert Systems with Applications*, vol. 279, p. 127458, 2025.

[20] T. Schreiber, Measuring information transfer, *Physical Review Letter*, vol. 85, pp. 461–464, Jul 2000.

[21] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, The graph neural network model, *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[22] A. Vaswani, Attention is all you need, *Advances in Neural Information Processing Systems*, 2017.

[23] Y. Liu, H. Wu, J. Wang, and M. Long, Non-stationary transformers: Exploring the stationarity in time series forecasting, *Advances in Neural Information Processing Systems*, vol. 35, pp. 9881–9893, 2022.

[24] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 11106–11115, 2021.

[25] H. Wu, J. Xu, J. Wang, and M. Long, Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, *Advances in Neural Information Processing Systems*, vol. 34, pp. 22419–22430, 2021.

[26] A. Graves and A. Graves, Long short-term memory, *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] B. Yu, H. Yin, and Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 7 2018.

[29] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, Graph wavenet for deep spatial-temporal graph modeling, in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 7 2019.

[30] L. Chen, L. Gu, and Y. Fu, When semantic segmentation meets frequency aliasing, in *International Conference on Representation Learning*, vol. 2024, pp. 43194–43216, 2024.

[31] L. Chen, Y. Fu, L. Gu, D. Zheng, and J. Dai, Spatial frequency modulation for semantic segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 11, pp. 9767–9784, 2025.

[32] L. Chen, L. Gu, D. Zheng, and Y. Fu, Frequency-adaptive dilated convolution for semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3414–3425, June 2024.

[33] L. Chen, L. Gu, L. Li, C. Yan, and Y. Fu, Frequency dynamic convolution for dense image prediction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30178–30188, June 2025.

[34] L. Chen, L. Gu, and Y. Fu, Frequency-dynamic attention modulation for dense prediction, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22620–22632, October 2025.

[35] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, Frequency-aware feature fusion for dense image prediction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10763–10780, 2024.

Author biography

Ke Wang received the B.E. degree in Automation from the University of Electronic Science and Technology of China in 2024. He is currently pursuing the M.E. degree in Electronic and Information Engineering at the Southeast University–Monash University Joint Graduate School in Suzhou, China. His research interests include deep learning, photovoltaic power forecasting, and time series prediction.

Shibo Wang received the B.E. and Ph.D. degrees in Electrical Engineering from Shandong University, Jinan, China, in 2009 and 2015, respectively. He is currently a professor-level senior engineer and recognized as a leading technical expert in Jinan, China. His research interests include grid integration and operation of large-scale renewable energy and energy storage systems, control of distributed generation and microgrids, safety assessment

of renewable energy grid connection, field testing, technical supervision, and standard development.

Ju Wu received the Bachelor of Engineering (B.E.) degree in Wind Energy and Power Engineering from North China Electric Power University (Beijing), Beijing, China, in 2014, and the Master of Engineering (M.E.) degree in Renewable Energy and Clean Energy from the same university in 2017. He is currently an engineer at the China Electric Power Research Institute in Nanjing, China. His research interest is new energy power rate prediction technology.

Jingxin Zhang received B.E. degree in School of Electrical Engineering and Automation from Harbin Engineering University, Harbin, China, the M.E. degree in Control Science and Engineering from Harbin Institute of Technology, Harbin, China, in 2014 and 2016, respectively, and the Ph.D. degree in Control Science and Engineering from Tsinghua University, Beijing, China, in 2022. She is currently a lecturer with the Department of Automation, Southeast University in Nanjing, China. Her research interests are continual learning, data-driven fault detection and diagnosis, performance monitoring, photovoltaic power prediction and their applications in the industrial processes.