applied sciences

MDPI

Article

Dual-Encoder Transformer for Short-Term Photovoltaic Power Prediction Using Satellite Remote-Sensing Data

Haizhou Cao [1,2], Jing Yang [3], Xuemeng Zhao [3], Tiechui Yao [1,2], Jue Wang [1,2,*], Hui He [3] and Yangang Wang [1,2]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
* Correspondence: wangjue@sccas.cn

Abstract: The penetration of photovoltaic (PV) energy has gained a significant increase in recent years because of its sustainable and clean characteristics. However, the uncertainty of PV power affected by variable weather poses challenges to an accurate short-term prediction, which is crucial for reliable power system operation. Existing methods focus on coupling satellite images with ground measurements to extract features using deep neural networks. However, a flexible predictive framework capable of handling these two data structures is still not well developed. The spatial and temporal features are merely concatenated and passed to the following layer of a neural network, which is incapable of utilizing the correlation between them. Therefore, we propose a novel dual-encoder transformer (DualET) for short-term PV power prediction. The dual encoders contain wavelet transform and series decomposition blocks to extract informative features from image and sequence data, respectively. Moreover, we propose a cross-domain attention module to learn the correlation between the temporal features and cloud information and modify the attention modules with the spare form and Fourier transform to improve their performance. The experiments on real-world datasets, including PV station data and satellite images, show that our model achieves better results than other models for short-term PV power prediction.

Keywords: transformer; photovoltaic power forecasting; satellite images; deep learning

check for updates

Citation: Cao, H.; Yang, J.; Zhao, X.; Yao, T.; Wang, J.; He, H.; Wang, Y. Dual-Encoder Transformer for Short-Term Photovoltaic Power Prediction Using Satellite Remote-Sensing Data. *Appl. Sci.* **2023**, *13*, 1908. https://doi.org/10.3390/app13031908

Academic Editor: Sergio Toscani

Received: 10 December 2022
Revised: 19 January 2023
Accepted: 31 January 2023
Published: 1 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Facing the depletion of finite fossil fuels and the demands of carbon emission reduction, photovoltaic (PV) energy has raised widespread interest in recent years [1,2]. As a sustainable and flexible distributed energy source, PV power has been a significant part of the power grid with the rapid growth of its installed capacity worldwide [3]. However, the high intermittency and uncertainty of PV power make accurate prediction difficult and further pose technical challenges for reliable power system operation and control [4,5]. Meteorological factors, e.g., solar irradiance, temperature, wind speed and direction, cloud changes, and air pressure, are the primary causes of this arbitrary fluctuation. Therefore, effective PV prediction places critical demands on the model's capacity to take into account and handle various meteorological factors.

For PV power prediction, the prediction model and the forecast horizon mainly depend on the data sources. Traditional methods such as ARIMA [6] and exponential smoothing [7] hardly deal with multivariate data and thus usually take historical PV power series as the sole input. However, they are still limited to modeling non-stationary changes. To leverage the meteorological information, many previous studies mainly concentrate on employing local measurements [8] or numerical weather prediction (NWP) data [9]. However, these methods are incapable of capturing the volatility of PV power affected by cloud changes due to ignoring the data sources containing cloud conditions. More specifically, the solar irradiance reaching the PV panels mainly depends on the amount and

motion of clouds [10,11] recorded in consecutive cloud images. In light of this, ground-based or satellite remote-sensing images are valuable data sources to support accurate short-term PV power prediction.

Several approaches employ ground-based/sky images for the prediction of irradiance [12,13] and PV power [14]. With sky images, the cloud motion vector (CMV) methods that can calculate the speed and direction of clouds have been developed to improve the irradiance prediction accuracy [15,16]. These methods have a considerably decreasing value beyond a 30 min future horizon and are thus commonly used in ultra-short-term PV power prediction [17]. Moreover, ground-based images are limited in observation range and require equipment installation. In contrast, satellite images offer top-down detection with greater spatial coverage, as well as advantages for prediction up to a few hours ahead [18]. They are not only suitable for CMV methods [11,19–21] but also able to provide other valuable information. Cai et al. [22] cluster the interval gray value of satellite images to construct a relationship with PV power. Wang et al. [23] propose a CNN-based fluctuation pattern prediction model with satellite images as input and apply LSTMs to individually forecast different patterns of PV power. Besides these methods that adopt only satellite images, researchers are increasingly interested in coupling satellite data with ground measurements to improve forecasting performance [24]. Si et al. [25] extracted cloud cover factor from satellite images by the modified CNN and then combined it with meteorological information for irradiance prediction. Agoua et al. [26] presented a spatiotemporal model with Lasso to integrate multi-source data for 6-h-ahead prediction. Furthermore, they used the proposed model to assess the impact of each source on forecasting performance, which showed that satellite images improved the accuracy by 3%. Yao et al. [27] proposed an encoder–decoder model, which includes U-Net and LSTM to extract the spatial features from satellite short-wave radiation data and temporal features from meteorological sequences, respectively. Then, the spatial and temporal features were concatenated for intra-hour PV power prediction.

However, a flexible predictive framework capable of handling data sources with different structures (i.e., image and sequence) is still not well discussed. Due to the insufficient consideration of the characteristics of these two structures, the spatiotemporal features extracted using the above methods are still not informative for PV power prediction up to hours. Although various deep neural networks have been applied in PV power prediction, the majority of existing methods consist of conventional deep learning models such as CNNs, RNNs, and fully connected layers. At present, few studies have followed the state-of-the-art transformer [28] architecture, which has achieved great success in natural language processing [29] and computer vision [30]. Compared with convolutional or recurrent neural networks, transformer-based models with the self-attention mechanism are superior in capturing temporal dependencies. Simeunovic et al. [31] combined a graph neural network with a transformer to propose a graph–convolutional transformer for multi-site PV power prediction based on historical weather data. In addition, most existing models based on hybrid neural networks merely concatenate the spatial and temporal features before passing them as a whole input to the following layer. As a result, this rough treatment is incapable of learning and utilizing the correlation between the features of structured local measurements and unstructured satellite images well.

Motivated by the aforementioned issues, we renovate the transformer architecture and propose a novel transformer-based PV power prediction model that can process the locally measured sequences and satellite remote-sensing images. The main contributions are summarized as follows:

1. We propose a novel **dual-e**ncoder **t**ransformer (DualET) for short-term PV power prediction. Distinct from the standard transformer architecture, DualET contains dual encoders, including a local seasonal information (LSI) encoder and a remote-sensing information (RSI) encoder, and a single decoder. The dual encoders are designed to handle sequence and image data, while the decoder is to model the joint meteorological features from the encoders and outputs the short-term prediction.

2. To extract informative features from the satellite images and local sequences, we deploy the two-dimensional wavelet transform block and series decomposition block in the encoding stage. The former in the RSI encoder is capable of the frequency feature extraction of image signals to obtain cloud detailed information, while the latter in the LSI encoder conducts series decomposition to improve the learning capacity of temporal patterns.
3. We propose a cross-domain attention module to learn the correlation between the temporal features in sequence data and the cloud information in image data. Furthermore, we enhance the ability to capture dependencies by modifying the attention modules with the sparse form and Fourier transform.
4. Real-world datasets are applied to evaluate the proposed model. The experiments show that our model achieves state-of-the-art results compared with other models including recent transformers.

2. Problem Definition

Two types of data sources are used in this study to predict the PV power, including cloud images processed from the satellite remote sensing data and in situ data from real PV stations. The in situ data contain historical records of several meteorological factors and PV power. The detailed data description is provided in Section 4.1. We denote the historical steps of input data as L_{in} and the prediction steps as L_{out}. The number of in situ data attributes and remote-sensing data channels used is denoted as D_{LS} and C_{RS}, respectively. In summary, this PV power forecasting problem can be formulated as

$$f(\mathbf{X}_{\text{RS}}; \mathbf{X}_{\text{LS}}) \longrightarrow \hat{\mathbf{y}} \tag{1}$$

where $f(\cdot)$ is the mapping function, i.e., the forecasting model; $\mathbf{X}_{\text{RS}} \in \mathbb{R}^{L_{\text{in}} \times H \times W \times C_{\text{RS}}}$, $\mathbf{X}_{\text{LS}} \in \mathbb{R}^{L_{\text{in}} \times D_{\text{LS}}}$ are the input remote-sensing data and local sequences (in situ data), respectively; $\hat{\mathbf{y}} \in \mathbb{R}^{L_{\text{out}}}$ is the PV power prediction. The hidden dimension of the model is denoted as D.

3. Methodology

PV power generation substantially fluctuates with meteorological factors, such as solar irradiance, temperature, wind speed/direction, and cloud changes. For an accurate PV power prediction, we propose a novel dual-encoder transformer (DualET) to capture context features from these factors. As shown in Figure 1, DualET has an encoder–decoder architecture like most transformers but with dual encoders. One encoder, i.e., local seasonal information encoder (LSI encoder), is to model temporal dynamics from in situ measurements. The other, i.e., remote-sensing information encoder (RSI encoder), is to learn the spatial and temporal features of clouds from satellite images. The output of the LSI encoder and that of the RSI encoder contain fluctuations in the local measurements and clouds, respectively. Both of them can be combined to provide comprehensive fluctuation information for modeling the changes in the PV power series. Further, we design a joint feature decoder to predict future short-term PV power. The details of DualET will be introduced in the following subsections.

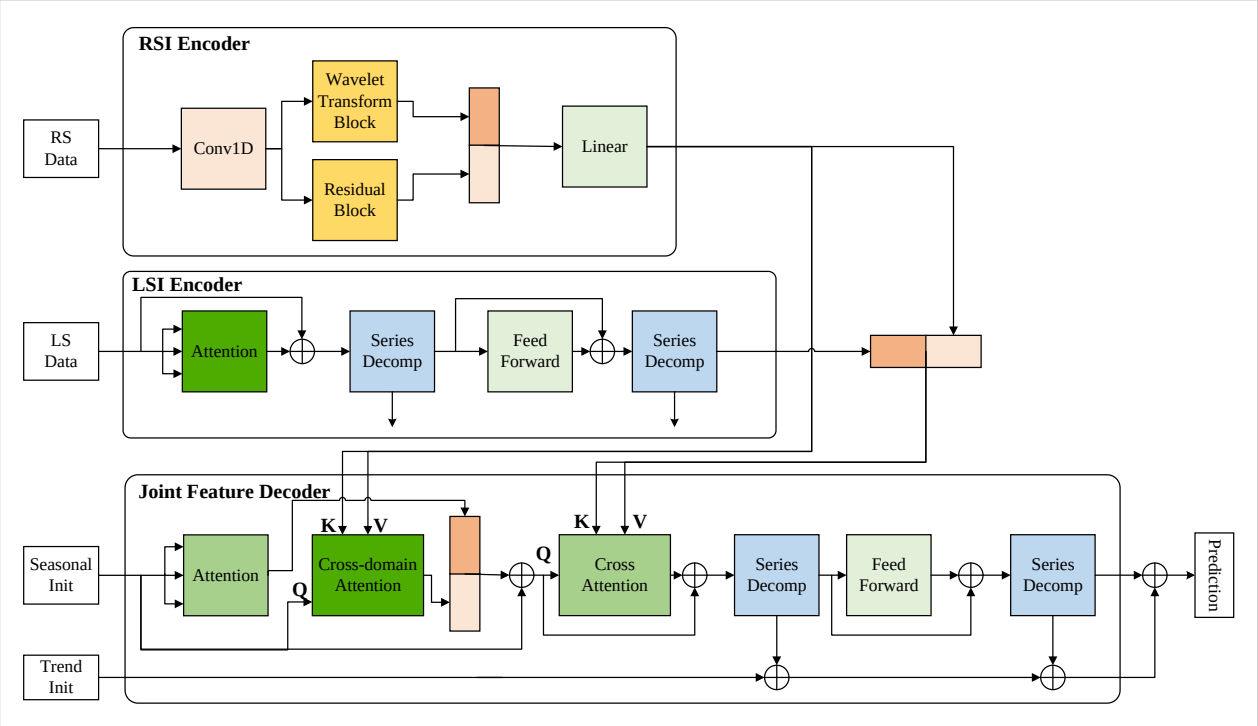

Figure 1. Model architecture of DualET.

3.1. Decomposition Modules

3.1.1. Wavelet Transform Block

The object edges represent abrupt changes around smooth regions and are distributed in the high-frequency signals of an image. For the motion and coverage of clouds, cloud edges offer crucial information. Wavelet transform is a powerful analysis tool widely used in image signal processing [32]. Compared with Fourier transform, it can capture frequency properties without location information loss. Therefore, we build the wavelet transform block (WTBlock) to extract frequency features and provide the RSI encoder with additional cloud details. As shown in Figure 2, the image signals are passed through high-pass and low-pass filters, sequentially from the horizontal and vertical directions. The high-pass filter (HPF) is to extract high-frequency components such as edges, while the low-pass filter (LPF) is to obtain low-frequency components for approximation. We summarize this 2D discrete wavelet transform as

$$\mathbf{I}_{\mathrm{LL}}, \mathbf{I}_{\mathrm{LH}}, \mathbf{I}_{\mathrm{HL}}, \mathbf{I}_{\mathrm{HH}} = \mathrm{WTBlock}(\mathbf{I}), \tag{2}$$

where image signal $\mathbf{I}$ is decomposed into four components: the approximation $\mathbf{I}_{\mathrm{LL}}$ (passing through LPFs in both directions), and three details ($\mathbf{I}_{\mathrm{LH}}, \mathbf{I}_{\mathrm{HL}}, \mathbf{I}_{\mathrm{HH}}$) in horizontal, vertical, and diagonal orientations, respectively.

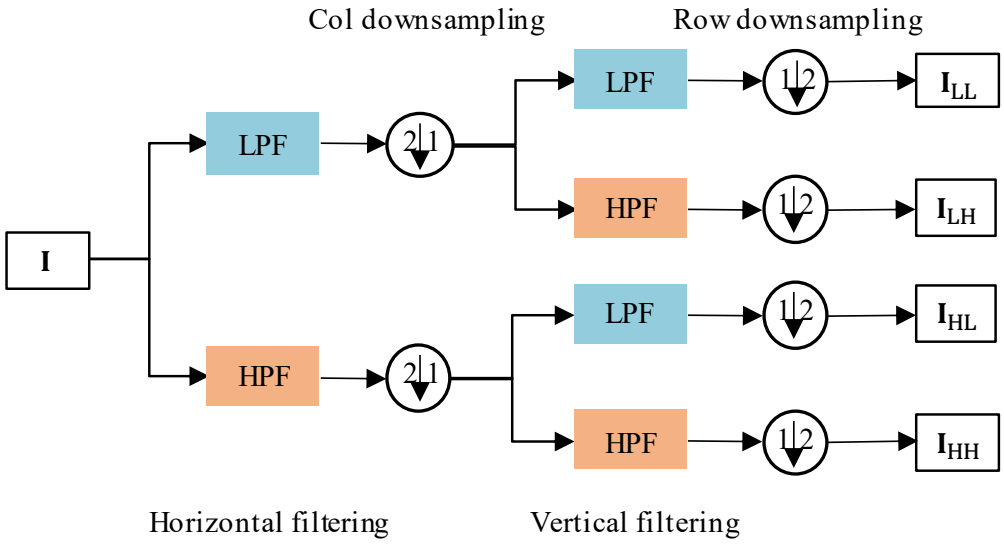

Figure 2. Illustration of 2D wavelet transform. LPF denotes the low-pass filter and HPF denotes the high-pass filter.

3.1.2. Series Decomposition Block

Real-world sequential signals (e.g., PV power sequence) are often entangled with temporal patterns that are informative for forecasting. Time series decomposition is an effective strategy to decouple knotted patterns from sequences. Among decomposition methods, seasonal-trend decomposition [33] has been widely employed as a feature engineering technique, which separates sequences into seasonal and trend parts. Inspired by Autoformer [34], we apply this decomposition idea as a series decomposition block (SDBlock) to enhance the pattern extraction ability of DualET. Given an input sequence $\mathbf{X}$, the procedure is

$$\mathbf{S}, \mathbf{T} = \text{SDBlock}(\mathbf{X}), \tag{3}$$

where $\mathbf{T}$ is the moving average result of $\mathbf{X}$ and is considered as the trend part; $\mathbf{S}$ is the residual part (i.e., the detrend part), which is regarded as the seasonal part. To keep the sequence length unchanged, a padding operation is performed on the input sequence.

3.2. Learning Modules

3.2.1. Residual Connection and Residual Block

The residual network architecture (i.e., residual connection) has become the foundation of deep neural networks, which learn residual mapping to ease the optimization of deep layers [35]. It can be formalized as $y = F(x) + x$, i.e., the input is added to the output of stacked layers (F) as the result. For the RSI encoder, we employ residual blocks to learn the representation of remote-sensing data. As shown in Figure 3, the residual block is stacked using two convolutions, batch normalization [36], and activation (ReLU) layers. The process is summarized as

$$\mathbf{X}_{\text{R}} = \text{ResBlock}(\mathbf{X}). \tag{4}$$

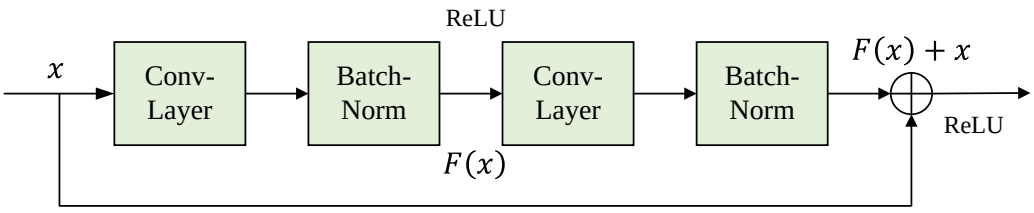

Figure 3. Illustration of residual block.

3.2.2. Attention Mechanism

As one of the most representative identifiers of transformers, the attention mechanism is proposed as a query–key–value (QKV) model to learn long-range dependencies without recurrent structures. Given the matrices $\mathbf{Q} \in \mathbb{R}^{L_q \times D_k}$, $\mathbf{K} \in \mathbb{R}^{L_k \times D_k}$, and $\mathbf{V} \in \mathbb{R}^{L_k \times D_v}$ as the projected queries, keys, and values, the single-head version of standard attention mechanism can be formalized as $\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_k}})\mathbf{V}$, where L_q, L_k denote the length of queries and keys/values; D_k, D_v denote the projected dimensions; $\frac{1}{\sqrt{D_k}}$ is the scale factor to avoid $\text{Softmax}(\cdot)$ yielding extremely small gradients. Furthermore, the multi-head version is as follows:

$$\begin{aligned} &\mathcal{A}_{\text{multi-head}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \cdots, \text{head}_H)\mathbf{W}^O, \\ &\text{where head}_i = \mathcal{A}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \end{aligned} \tag{5}$$

The queries, keys, and values with dimension D are mapped into H heads (i.e., subspaces) by $\mathbf{W}_i^Q, \mathbf{W}_i^K \in \mathbb{R}^{D \times D_k}$, and $\mathbf{V}\mathbf{W}_i^V \in \mathbb{R}^{D \times D_v}$. Then the outputs of these heads are concatenated and mapped back to D by $\mathbf{W}^O \in \mathbb{R}^{HD_v \times D}$. In most cases, $D_k = D_v = D/H$. The standard transformer has two types of multi-head attention: self-attention and cross-attention. For self-attention, its projected queries, keys, and values are shared with the same source, while the key–value pairs of cross-attention are typically from the output of encoder.

In practice, the attention modules used in DualET are modified to improve the performance of capturing dependencies. First, we design an additional cross-domain attention module to discover the correlation between the features of images and sequences. Concretely, the queries are the temporal features of the decoder, and the key–value pairs are the cloud information from the RSI encoder. Furthermore, we perform the fast Fourier transform (FFT) on the input and the inverse FFT on the output:

$$\mathcal{A}_{\text{FFT}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathcal{F}^{-1}(\mathcal{A}(\mathcal{F}(\mathbf{Q}), \mathcal{F}(\mathbf{K}), \mathcal{F}(\mathbf{V}))) \tag{6}$$

where $\mathcal{F}, \mathcal{F}^{-1}$ denote the FFT and its inverse, which are also used in the self-attention module of the LSI encoder. The FFT plays a key role in signal processing because it can rapidly convert a signal from the time/space domain to the frequency domain (and vice versa) and describe relationships between these domains [37]. It is defined by

$$\overline{\mathbf{X}}_k = \sum_{m=0}^{L-1} e^{-2\pi i k \cdot (m/L)} \mathbf{X}_m, \quad k = 0, \ldots, L-1 \tag{7}$$

Based on FFT, the attention module can discover the frequency correlation between queries and keys. In addition, we employ the ProbSparse attention mechanism [38] as the self-attention and cross-attention modules of the decoder to improve their performances.

3.2.3. Embedding and Feed-Forward Layer

For sequence modeling, the order information of time steps is crucial. Furthermore, the timestamp records of local sequences (meteorological and PV power series) are instructive for PV power prediction but are hardly utilized in the standard transformer architecture. To introduce this information, we employ timestamp-embedding layers (following Autoformer [34]) for the local sequence inputs.

The feed-forward layer is a position-wise fully connected module, i.e., the learnable parameters are shared with each step. It contains two linear layers ($\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2$) and a ReLU activation function in between, formulated as

$$\text{FeedFoward}(\mathbf{X}) = \text{ReLU}(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2. \tag{8}$$

3.3. Remote-Sensing Information Encoder

The RSI encoder is designed to learn spatial and temporal features from remote-sensing data. As shown in the top diagram of Figure 1, it is mainly composed of a two-dimensional convolution layer, a wavelet transform block, and a residual block. Given the historical L_{in} steps of cloud images $\mathbf{X}_{\text{RS}} \in \mathbb{R}^{L_{\text{in}} \times H \times W \times C_{\text{RS}}}$ as the input of the RSI encoder, the procedure is

$$\begin{aligned}
&\mathbf{I}_1 = \text{Conv2D}(\mathbf{X}_{\text{RS}}),\\
&\mathbf{I}_{\text{LL}}, \mathbf{I}_{\text{LH}}, \mathbf{I}_{\text{HL}}, \mathbf{I}_{\text{HH}} = \text{WTBlock}(\mathbf{I}_1),\\
&\mathbf{I}_2 = \text{Conv2D}(\text{ResBlock}(\mathbf{I}_1)),\\
&\mathbf{I}_3 = \text{ConvFusion}(\text{Concat}([\mathbf{I}_2, \mathbf{I}_{\text{LL}}, \mathbf{I}_{\text{LH}} + \mathbf{I}_{\text{HL}} + \mathbf{I}_{\text{HH}}])),\\
&\mathbf{Z}_{\text{RS}} = \text{Linear}(\text{Flatten}(\text{ReLU}(\mathbf{I}_3))),
\end{aligned} \tag{9}$$

where ConvFusion is 1×1 2D convolution to integrate frequency components with image features; $\mathbf{Z}_{\text{RS}} \in \mathbb{R}^{L_{\text{in}} \times D}$ is the output of the RSI encoder.

3.4. Local Seasonal Information Encoder

For weather and PV power series, the seasonal part contains the main volatility feature, which is the key to accurate prediction. Therefore, we introduce series decomposition blocks to the LSI encoder to extract seasonal patterns from local measurement data. As shown in the middle of Figure 1, the LSI encoder is staked with L_{LSI} LSI encoder layers. Given the input of the LSI encoder $\mathbf{X}^0_{\text{LS}} \in \mathbb{R}^{L_{\text{in}} \times D}$ that is embedded from $\mathbf{X}_{\text{LS}} \in \mathbb{R}^{L_{\text{in}} \times D_{\text{DS}}}$, the procedure in l-th LSI encoder layer is

$$\begin{aligned}
&\mathbf{S}^l_{\text{LS},1}, _ = \text{SDBlock}(\text{Attention}(\mathbf{X}^{l-1}_{\text{LS}}) + \mathbf{X}^{l-1}_{\text{LS}}),\\
&\mathbf{S}^l_{\text{LS},2}, _ = \text{SDBlock}(\text{FeedFoward}(\mathbf{S}^l_{\text{LS},1}) + \mathbf{S}^l_{\text{LS},1}),\\
&\mathbf{X}^l_{\text{LS}} = \mathbf{S}^l_{\text{LS},2},
\end{aligned} \tag{10}$$

where "_" is the ignored trend part; $\mathbf{S}^l_{\text{LS,i}}, i \in \{1, 2\}$ denotes the seasonal part in the l-th layer; $\mathbf{Z}_{\text{LS}} = \mathbf{X}^{L_{\text{LSI}}}_{\text{LS}}$ denotes the output of the LSI encoder.

3.5. Joint-Feature Decoder

The joint-feature decoder is to model temporal dynamics based on the joint features of local and remote-sensing data and then outputs short-term PV power. As shown in the bottom diagram of Figure 1, the decoder is stacked with L_{de} decoder layers, and each layer contains three attention modules (i.e., self-attention, cross-domain attention, and cross-attention) to determine the correlations from different perspectives. The outputs of two encoders are integrated as the input of cross-attention by ConvFusion: $\mathbf{Z}_{\text{en}} = \text{ConvFusion}(\text{Concat}([\mathbf{Z}_{\text{RS}}, \mathbf{Z}_{\text{LS}}]))$. The inputs of the decoder contain initialized seasonal part $\mathbf{X}^0_{\text{de}}$ and trend part $\mathbf{T}^0_{\text{de}}$. which are decomposed from the latter half of $\mathbf{X}^0_{\text{LS}}$ and concatenated with scalar placeholders (zeros for the seasonal part and means for the trend part). The details in the l-th decoder layer are

$$\begin{aligned}
&\mathbf{Z}^l_{\text{de},1} = \text{Attention}(\mathbf{X}^{l-1}_{\text{de}}),\\
&\mathbf{Z}^l_{\text{de},2} = \text{CrossDomianAttention}(\mathbf{X}^{l-1}_{\text{de}}, \mathbf{Z}_{\text{RS}}),\\
&\mathbf{Z}^l_{\text{de},3} = \mathbf{X}^{l-1}_{\text{de}} + \text{ConvFusion}(\text{Concat}([\mathbf{Z}^l_{\text{de},1}, \mathbf{Z}^l_{\text{de},2}])),\\
&\mathbf{S}^l_{\text{de},1}, \mathbf{T}^l_{\text{de},1} = \text{SDBlock}(\text{Attention}(\mathbf{Z}^l_{\text{de},3}, \mathbf{Z}_{\text{en}}) + \mathbf{Z}^l_{\text{de},3}),\\
&\mathbf{X}^l_{\text{de}}, \mathbf{T}^l_{\text{de},2} = \text{SDBlock}(\text{FeedFoward}(\mathbf{S}^l_{\text{de},1}) + \mathbf{S}^l_{\text{de},1}),\\
&\mathbf{T}^l_{\text{de}} = \mathbf{T}^{l-1}_{\text{de}} + \mathcal{W}^l_1 * \mathbf{T}^l_{\text{de},1} + \mathcal{W}^l_2 * \mathbf{T}^l_{\text{de},2},
\end{aligned} \tag{11}$$

where $\mathbf{Z}^l_{\text{de,i}}, i \in \{1, 2, 3\}$ is the intermediate feature; $\mathbf{S}^l_{\text{de},1}, \mathbf{X}^l_{\text{de}}$ denote the seasonal parts in l-th layer; $\mathbf{T}^l_{\text{de},1} (i \in \{1, 2\}), \mathbf{T}^l_{\text{de}}$ denote the trend parts in l-th layer; $\mathcal{W}^l_i \in \{1, 2\})$ denote

the project functions for the trend parts. After L_{de} decoder layers, the final prediction $\hat{\mathbf{y}}$ is from the sum of two parts: $\mathcal{W} * \mathbf{X}_{\text{de}}^{L_{\text{de}}} + \mathbf{T}_{\text{de}}^{L_{\text{de}}}$, where $\mathcal{W}$ is the projector for the seasonal part.

4. Experiment

In this section, we evaluate the proposed DualET on satellite images and actual PV station data. We first introduce the datasets and data preprocessing. Then, we describe the experimental setting in detail. Finally, we compare the prediction performance of DualET and the baseline models and conduct several ablation experiments.

4.1. Datasets and Data Preprocessing

Two datasets were used in this study, namely, satellite remote-sensing data and PV station data. The satellite data were the L1 grid data from Himawa-8, a geostationary satellite launched in 2015 by the Japan Meteorological Agency to provide weather forecasts and typhoon and storm reports for Japan, East Asia, and the Western Pacific. The detection range of Himawa-8 is 60° N, 160° E, and 80° W, with a spatial resolution of 0.05°, which corresponds to about 5 km on the ground, and the temporal resolution is 10 min. The PV station data contains local measurements from three real PV stations at different latitudes and longitudes in Hebei, China. Each station records meteorological factors (including global and diffuse irradiance, temperature, wind direction and speed, and air pressure) and PV power records at 15 min intervals.

We set the temporal resolution as 30 min to harmonize the time intervals of these two datasets. The satellite remote-sensing data were processed into 40×40 cloud images centered on the latitude and longitude of the PV station. The satellite data were sampled from July 2018 to June 2019 to align with that of PV station data. For each day, the data with a time range from 7:00 to 19:00 (UTC +0800) were used. We divided these two datasets into a training set, a validation set, and a test set in the ratio of 8:1:1 after arranging them to ensure the test period covers multiple seasons. Before input to the model, the data were normalized using standardization to eliminate the inconsistency of the magnitude of each dimension.

4.2. Experimental Setting

4.2.1. Baseline Models

We primarily selected seven models as baselines for comparison, namely, a classic statistical model ARIMA [39], an RNN-based model LSTM [40], and three state-of-the-art models, i.e., Transformer [28] and its variants Informer [38] and Autoformer [34].

4.2.2. Hyperparameters and Platform

Our model DualET and transformer baselines, i.e., Transformer, Informer, and Autoformer, were set to the same number of layers, including two encoder layers and one decoder layer. The hidden dimension of model D was set to 512. The number of attention heads was set to 8. The batch size was set to 32, and the training epochs were set to 10 (with early stopping). The loss function was the mean-squared error (MSE) (Equation (12)), and the optimizer was ADAM with an initial learning rate of 1×10^{-4}. The input length of the model was set to 24, i.e., 12 h, and the output prediction length was set to 12, i.e., 6 h. All the models were implemented with PyTorch and conducted on a Ubuntu server with four NVIDIA GeForce RTX 2080Ti 11GB GPUs.

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2. \tag{12}$$

4.2.3. Evaluation Metrics

We evaluated the performance of the model with three widely used metrics, i.e., mean absolute error (MAE), root-mean-squared error (RMSE), and symmetric mean absolute percentage error (SMAPE).

$$\begin{aligned} \text{MAE} &= \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t|, \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \\ \text{SMAPE} &= \frac{100\%}{N}\sum_{t=1}^{N}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2}. \end{aligned} \tag{13}$$

4.3. Results

As shown in Table 1, we evaluated the prediction performance of the proposed DualET on three different PV stations. For short-term (6-h) PV power prediction, DualET achieved the best results on three error metrics: MAE, RMSE, and SMAPE. Compared with the results of other models on all stations, DualET achieved a relative MAE and RMSE reduction of 22.53% and 22.75%, respectively, which is a significant improvement. The average reduction in MAE was even more than 53%, compared with the traditional ARIMA, and 27.72% compared with the popular LSTM. Among these baselines, the transformer-based models, i.e., Transformer, Informer, and Autoformer, were better than ARIMA and LSTM models. Moreover, our DualET still outperformed the competitive transformer-based models and yielded a relative MAE reduction of 10.64%.

Table 1. Prediction performance of the proposed DualET.

Models	Metrics	Station0	Station1	Station2
ARIMA	MAE	3.6934	3.7727	3.6343
	RMSE	5.3633	6.3809	5.4976
	SMAPE	0.8881	0.9019	0.9135
LSTM	MAE	2.3500	2.2612	2.6287
	RMSE	3.0866	2.9739	3.4570
	SMAPE	0.7218	0.6888	0.7635
Transformer	MAE	2.0780	2.0678	1.8413
	RMSE	2.8468	2.7659	2.5875
	SMAPE	0.7110	0.6482	0.6274
Informer	MAE	1.8274	1.8393	1.8564
	RMSE	2.5874	2.5773	2.5940
	SMAPE	0.6549	0.6232	0.6498
Autoformer	MAE	1.9989	2.0690	1.9771
	RMSE	2.7379	2.7834	2.7700
	SMAPE	0.6905	0.6764	0.6810
DualET	MAE	**1.6904**	**1.7576**	**1.7657**
	RMSE	**2.2845**	**2.4205**	**2.5087**
	SMAPE	**0.6335**	**0.5931**	**0.6258**

The prediction results of different models are presented in Figures 4 and 5. It obviously shows that the number of deviation points predicted by the statistical model ARIMA is much larger than those predicted by other DNN-based models, which indicates that ARIMA is unsuitable for short-term prediction up to several hours, especially for the nonstationary PV power series. As shown in Figure 5, the LSTM model has a more scattered distribution of points than the transformer-based models, which indicates the significance of transformer

architecture for sequence modeling. It can also be seen that the proposed DualET presents the fittest curves in Figure 4 and the narrowest band in Figure 5 compared with other baselines, which shows the advantages of PV power prediction.

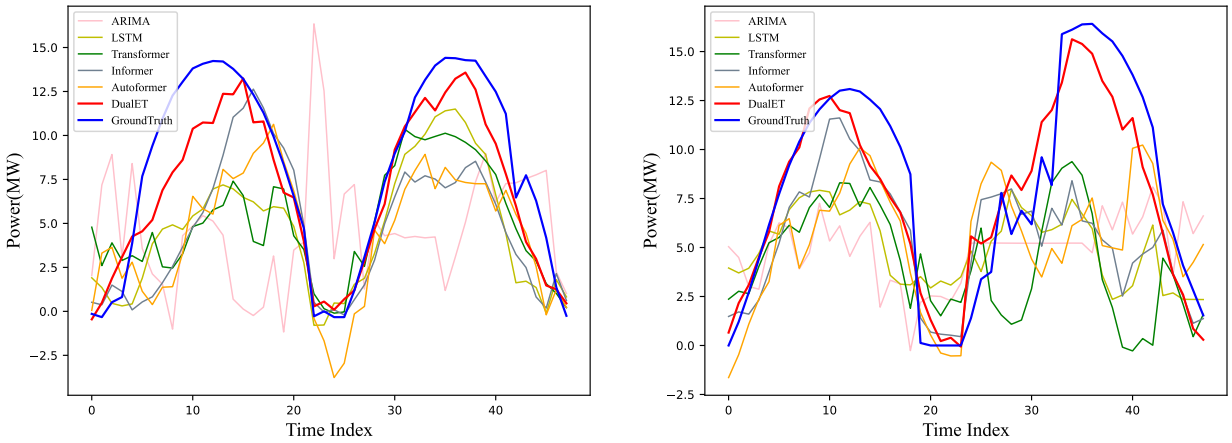

Figure 4. Prediction results of different models.

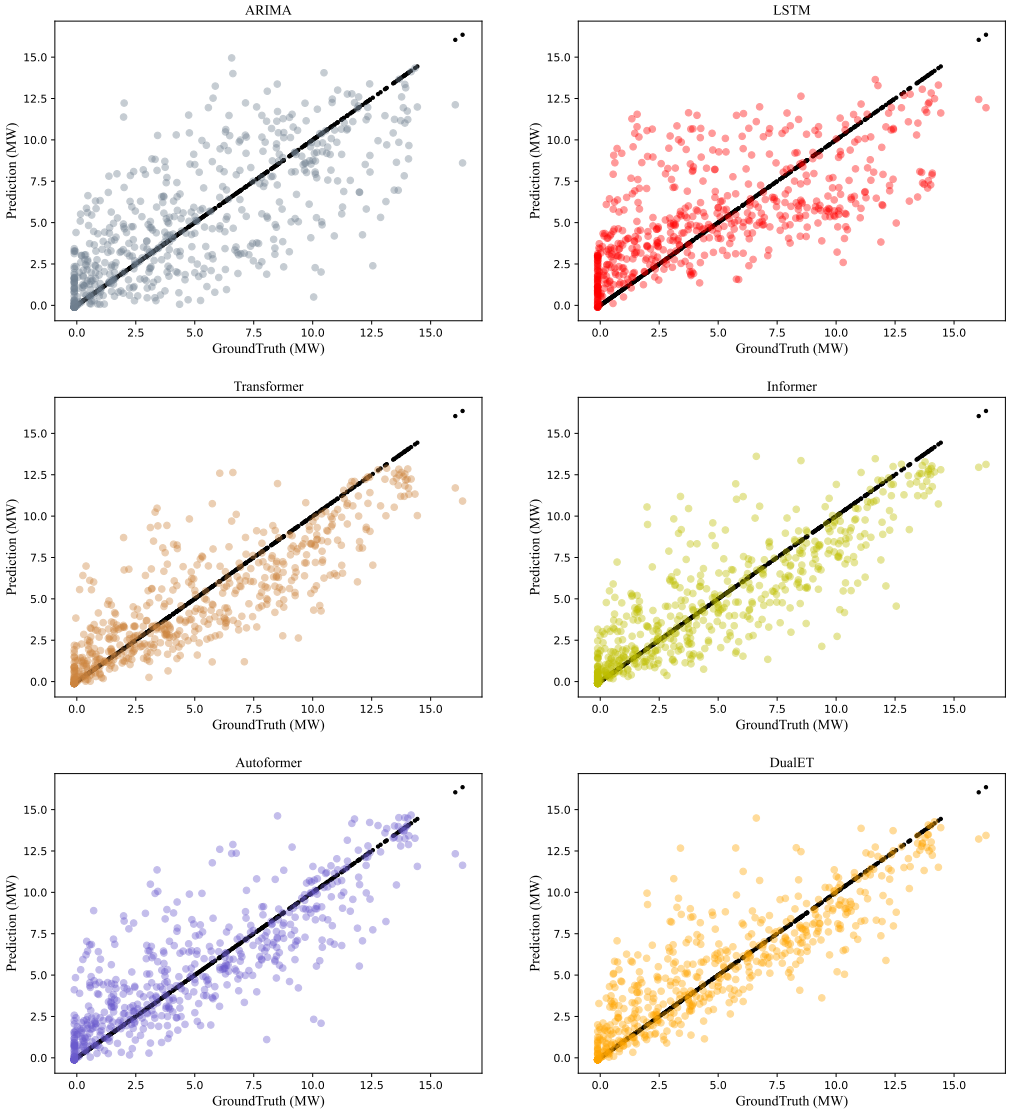

Figure 5. Scatter plots of different models.

4.4. Ablation Studies

DualET contains dual encoders, including the LSI encoder to deal with local seasonal information and the RSI encoder to process remote-sensing information, with a shared decoder to combine joint features. In addition, there are different decomposition modules and attention modules employed in DualET to enhance feature extraction. We conducted additional experiments to evaluate the impact of the dual encoders and different functional modules with MAE, RMSE, and SMAPE as evaluation metrics, and we present the results of these experiments in this section.

4.4.1. Different Encoders

We evaluated the impact of different encoders by keeping only one of the encoders from DualET. As shown in Table 2, the proposed model with only the LSI encoder is better than that with only the RSI encoder. As previously mentioned, the LSI encoder is used to learn historical seasonal information from local measurements, while the RSI encoder is to learn cloud information from satellite images. Therefore, this indicates that the fluctuation features of PV power are more contained in local measurements compared with cloud images. Meanwhile, we can see that the model with dual encoders yields the best results compared with that with only one, which shows that cloud images can provide valuable information for performance improvement and further shows the effectiveness of our dual-encoder design.

Table 2. Comparison of different encoders.

Models	MAE	RMSE	SMAPE
Only LSI encoder	1.9948	2.6964	0.6784
Only RSI encoder	2.0275	2.7453	0.6937
DualET	**1.6904**	**2.2845**	**0.6335**

4.4.2. Decomposition Blocks in Dual Encoders

We evaluated the impact of decomposition blocks for dual encoders by removing the series decomposition (SD) blocks or wavelet transform (WT) blocks. As shown in Table 3, the dual encoders without decomposition blocks will give some performance degradation. This shows that the WT and SD can enhance the learning capacity of dual encoders and provide informative features for the decoder. In addition, the performance of the RSI encoder without WT is worse than the LSI encoder without SD, which indicates that the high- and low-frequency features of cloud images are essential for short-term PV power prediction.

Table 3. Ablation studies of decomposition structures. SD and WT mean the series decomposition block and wavelet transform block, respectively. W/o means "without".

Models	MAE	RMSE	SMAPE
LSI w/o SD	1.9378	2.5927	0.6935
RSI w/o WT	1.9916	2.6440	0.7094
DualET	**1.6904**	**2.2845**	**0.6335**

4.4.3. Different Attention Modules

In DualET, the decoder's self- and cross-attention modules are formed with ProbSparse, while the others are enhanced by fast Fourier transform (FFT). As shown in Table 4, we evaluated the impact of different attention modules including the cross-domain attention module. We can see that if all attention modules are the same type (ProbSparse form or FFT), it results in performance degradation. Therefore, we employed the various attention modules with suitable modifications for better correlation capture. Moreover, the prediction error of the model without cross-domain attention significantly increases. This

demonstrates the effectiveness of cross-domain attention to learn the correlation between features in sequence data and image data.

Table 4. Different attention modules. W/o means "without".

Models	MAE	RMSE	SMAPE
W/o cross-domain attention	2.0192	2.7100	1.0466
All attention with FFT	1.8060	2.4965	0.6630
All attention with ProbSparse	1.7848	2.4507	0.6657
DualET	**1.6904**	**2.2845**	**0.6335**

5. Conclusions

To handle satellite images and ground measurements, in this paper, we proposed a novel transformer-based model with dual encoders, named DualET, for short-term PV power prediction. To obtain cloud detailed information from satellite images, a two-dimensional wavelet transform block and a residual block were used in the remote-sensing information encoder. For the local seasonal information encoder, we conducted self-attention and series decomposition to learn the temporal patterns from local sequences. For the decoder, we employed three types of attention modules and series decomposition blocks to model the joint features of local and remote-sensing information and output the prediction. Specifically, a cross-domain attention module was proposed to learn the correlation between the temporal features and cloud information. Finally, the experiments on real-world datasets, including PV station data and satellite images, were presented to show the prediction performance of DualET. In addition, the ablation studies show the effectiveness of our design. In the future, we will attempt to improve the model architecture so that more data sources (e.g., NWP or other satellite remote sensing data) can be utilized to predict a longer horizon.

Author Contributions: Conceptualization, J.W. and H.H.; methodology, H.C. and J.Y.; software, H.C. and J.Y.; validation, X.Z., T.Y. and Y.W.; writing—original draft preparation, H.C. and X.Z.; writing—review and editing, T.Y., J.W. and Y.W.; visualization, J.Y. and X.Z.; supervision, J.W. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (2021ZD0110403).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The remote-sensing data used in this paper were processed from Himawari-8 satellite data supplied by the P-Tree System, Japan Aerospace Exploration Agency (JAXA). We also gratefully acknowledge the support of the MindSpore team.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Kabir, E.; Kumar, P.; Kumar, S.; Adelodun, A.A.; Kim, K.H. Solar Energy: Potential and Future Prospects. *Renew. Sustain. Energy Rev.* **2018**, *82*, 894–900. [CrossRef]
2. Armeanu, D.S.; Joldes, C.C.; Gherghina, S.C.; Andrei, J.V. Understanding the Multidimensional Linkages among Renewable Energy, Pollution, Economic Growth and Urbanization in Contemporary Economies: Quantitative Assessments across Different Income Countries' Groups. *Renew. Sustain. Energy Rev.* **2021**, *142*, 110818. [CrossRef]
3. Carriere, T.; Vernay, C.; Pitaval, S.; Kariniotakis, G. A Novel Approach for Seamless Probabilistic Photovoltaic Power Forecasting Covering Multiple Time Frames. *IEEE Trans. Smart Grid* **2019**, *11*, 2281–2292. [CrossRef]
4. Sanjari, M.J.; Gooi, H. Probabilistic Forecast of PV Power Generation Based on Higher Order Markov Chain. *IEEE Trans. Power Syst.* **2016**, *32*, 2942–2952. [CrossRef]

5. Stein, G.; Letcher, T.M. Integration of PV Generated Electricity into National Grids. In *A Comprehensive Guide to Solar Energy Systems*; Letcher, T.M., Fthenakis, V.M., Eds.; Academic Press: Cambridge, MA, USA, 2018; pp. 321–332. [CrossRef]
6. Li, Y.; Su, Y.; Shu, L. An ARMAX Model for Forecasting the Power Output of a Grid Connected Photovoltaic System. *Renew. Energy* **2014**, *66*, 78–89. [CrossRef]
7. Prema, V.; Rao, K.U. Development of Statistical Time Series Models for Solar Power Prediction. *Renew. Energy* **2015**, *83*, 100–109. [CrossRef]
8. Liu, L.; Zhan, M.; Bai, Y. A Recursive Ensemble Model for Forecasting the Power Output of Photovoltaic Systems. *Sol. Energy* **2019**, *189*, 291–298. [CrossRef]
9. Böök, H.; Lindfors, A.V. Site-Specific Adjustment of a NWP-Based Photovoltaic Production Forecast. *Sol. Energy* **2020**, *211*, 779–788. [CrossRef]
10. Breitkreuz, H.; Schroedter-Homscheidt, M.; Holzer-Popp, T.; Dech, S. Short-Range Direct and Diffuse Irradiance Forecasts for Solar Energy Applications Based on Aerosol Chemical Transport and Numerical Weather Modeling. *J. Appl. Meteorol. Climatol.* **2009**, *48*, 1766–1779. [CrossRef]
11. Kato, T.; Manabe, Y.; Funabashi, T.; Yoshiura, K.; Kurimoto, M.; Suzuoki, Y. A Study on Several Hours Ahead Forecasting of Spatial Average Irradiance Using NWP Model and Satellite Infrared Image. In Proceedings of the 2016 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Beijing, China, 16–20 October 2016; pp. 1–8.
12. Zhang, C.; Du, Y.; Chen, X.; Lu, D.D.C. Cloud Motion Tracking System Using Low-Cost Sky Imager for PV Power Ramp-Rate Control. In Proceedings of the 2018 IEEE International Conference on Industrial Electronics for Sustainable Energy Systems (IESES), Hamilton, New Zealand, 31 January–2 February 2018; pp. 493–498.
13. Zhao, X.; Wei, H.; Wang, H.; Zhu, T.; Zhang, K. 3D-CNN-Based Feature Extraction of Ground-Based Cloud Images for Direct Normal Irradiance Prediction. *Sol. Energy* **2019**, *181*, 510–518. [CrossRef]
14. Lin, F.; Zhang, Y.; Wang, J. Recent Advances in Intra-Hour Solar Forecasting: A Review of Ground-Based Sky Image Methods. *Int. J. Forecast.* **2022**, *39*, 244–265. [CrossRef]
15. Bosch, J.L.; Kleissl, J. Cloud Motion Vectors from a Network of Ground Sensors in a Solar Power Plant. *Sol. Energy* **2013**, *95*, 13–20. [CrossRef]
16. Peng, Z.; Yu, D.; Huang, D.; Heiser, J.; Kalb, P. A Hybrid Approach to Estimate the Complex Motions of Clouds in Sky Images. *Sol. Energy* **2016**, *138*, 10–25. [CrossRef]
17. Tuohy, A.; Zack, J.; Haupt, S.E.; Sharp, J.; Ahlstrom, M.; Dise, S.; Grimit, E.; Mohrlen, C.; Lange, M.; Casado, M.G.; et al. Solar Forecasting: Methods, Challenges, and Performance. *IEEE Power Energy Mag.* **2015**, *13*, 50–59. . 2461351. [CrossRef]
18. Kallio-Myers, V.; Riihelä, A.; Lahtinen, P.; Lindfors, A. Global Horizontal Irradiance Forecast for Finland Based on Geostationary Weather Satellite Data. *Sol. Energy* **2020**, *198*, 68–80. [CrossRef]
19. Cros, S.; Liandrat, O.; Sebastien, N.; Schmutz, N. Extracting cloud motion vectors from satellite images for solar power forecasting. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 4123–4126.
20. Kebir, N.; Maaroufi, M. Best-effort algorithm for predicting cloud motion impact on solar PV power systems production. In Proceedings of the 2018 6th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), Istanbul, Turkey, 25–26 April 2018; pp. 34–38.
21. Zhou, Z.; Zhang, X.; Zhen, Z.; Mei, S. Cloud Displacement Vector Calculation in Satellite Images Based on Cloud Pixel Spatial Aggregation and Edge Matching for PV Power Forecasting. In Proceedings of the 2020 IEEE Sustainable Power and Energy Conference (iSPEC), Chengdu, China, 23–25 November 2020; pp. 112–119.
22. Cai, Y.; Liu, H.; Hu, P.; Fu, Z.; Wang, Y.; Zhang, D.; Ma, X.; Li, S. Ultra-short-term Photovoltaic Power Prediction Based on Elman Neural Network and Satellite Cloud Images. In Proceedings of the 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2), Taiyuan, China, 22–24 October 2021; pp. 2149–2154.
23. Wang, C.; Lu, X.; Zhen, Z.; Wang, F.; Xu, X.; Ren, H. Ultra-Short-Term Regional PV Power Forecasting Based on Fluctuation Pattern Recognition with Satellite Images. In Proceedings of the 2020 IEEE 3rd Student Conference on Electrical Machines and Systems (SCEMS), Jinan, China, 4–6 December 2020; pp. 970–975.
24. Blanc, P.; Remund, J.; Vallance, L. Short-term solar power forecasting based on satellite images. In *Renewable Energy Forecasting*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 179–198.
25. Si, Z.; Yu, Y.; Yang, M.; Li, P. Hybrid Solar Forecasting Method Using Satellite Visible Images and Modified Convolutional Neural Networks. *IEEE Trans. Ind. Appl.* **2020**, *57*, 5–16. [CrossRef]
26. Agoua, X.G.; Girard, R.; Kariniotakis, G. Photovoltaic Power Forecasting: Assessment of the Impact of Multiple Sources of Spatio-Temporal Data on Forecast Accuracy. *Energies* **2021**, *14*, 1432. [CrossRef]
27. Yao, T.; Wang, J.; Wu, H.; Zhang, P.; Li, S.; Xu, K.; Liu, X.; Chi, X. Intra-Hour Photovoltaic Generation Forecasting Based on Multi-Source Data and Deep Learning Methods. *IEEE Trans. Sustain. Energy* **2021**, *13*, 607–618. [CrossRef]
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.

29. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
30. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
31. Simeunovic, J.; Schubnel, B.; Alet, P.J.; Carrillo, R.E. Spatio-Temporal Graph Neural Networks for Multi-Site PV Power Forecasting. *IEEE Trans. Sustain. Energy* **2021**, *13*, 1210–1220. [CrossRef]
32. Walker, J.S. Wavelet-Based Image Processing. *Appl. Anal.* **2006**, *85*, 439–458. [CrossRef]
33. Robert, C.; William, C.; Irma, T. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J. Off. Stat.* **1990**, *6*, 3–73.
34. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 22419–22430.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 448–456.
37. Brigham, E.O. *The Fast Fourier Transform and Its Applications*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1988.
38. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]
39. Chen, P.; Pedersen, T.; Bak-Jensen, B.; Chen, Z. ARIMA-Based Time Series Model of Stochastic Wind Power Generation. *IEEE Trans. Power Syst.* **2009**, *25*, 667–676. [CrossRef]
40. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.