Contents list available at IJRED website

International Journal of Renewable Energy Development

Journal homepage: https://ijred.undip.ac.id

Research Article

Photovoltaic power prediction based on sky images and tokens-to-token vision transformer

Qiangsheng Dai[a*], Xuesong Hou[a], Dawei Su[a], Zhiwei Cui[b]

[a]*Power Dispatching and Control Center, State Grid Jiangsu Electric Power Co. Ltd., Nanjing, China*

[b]*Power Dispatching and Control Center, State Grid Taizhou Electric Power Co. Ltd, Taizhou, China*

Abstract. Photovoltaic (PV) power generation has high uncertainties due to the randomness and imbalance nature of solar energy and meteorological parameters. Hence, accurate PV power forecasts are essential in the operation of PV power plants (PVPP) for short-term dispatches and power generation schedules. In this paper, a new deep neural network structure based on vision transformer is proposed to combine sky images and Tokens-To-Token（T2T） for photovoltaic power prediction. The method uses an incremental tokenization module to aggregate neighboring image patches into tokens, which capture the local structural information of the clouds. Then, an efficient T2T-ViT backbone network is used to extract the global attentional relationships of the tokens for power prediction. In order to evaluate the performance of the proposed model, the method was compared with several deep learning architectures such as ResNet and GoogleNet on a dataset collected by the National Renewable Energy Laboratory in Colorado, USA. The results of power prediction were analysed using training loss, prediction error, and linear regression, and they show that the proposed method achieves higher prediction accuracy and lower error compared to the existing methods, especially in short- and ultra-short-term prediction. The paper demonstrates the potential of applying Transformer models to computer vision tasks for renewable energy forecasting. The results show that the proposed method achieves higher prediction accuracy and lower error than several deep learning architectures, such as ResNet and GoogleNet, especially in short- and ultra-short-term prediction.

Keywords: Photovoltaic power prediction, short term prediction, sky image, Deep learning, T2T-ViT

Received: 4th August 2023; Revised: 3rd Oct 2023; Accepted: 19th Oct 2023; Available online: 24th Oct 2023

1. Introduction

Renewable energy is deemed an alternative to fossil fuels for mitigating their detrimental environmental impact. Solar photovoltaic (PV) power generation has gained substantial attention recently due to its high reliability, low maintenance cost, quiet operation, and ability to generate environmentally friendly power. However, the variability of PV power output puts a serious burden on power system power regulation (Liu, Ren and Xu, 2021). Therefore, there is a growing need for accurate PV power prediction techniques, as reliable power prediction will contribute to efficient grid scheduling (Agoua, Girard, Kariniotakis, 2018).

Power prediction refers to the prediction of power output for a future period of time based on historical data and other relevant factors (Biswas *et al.*, 2021). In the past decades, many scholars have proposed many methods to forecast power, including the traditional Autoregressive Integrated Moving Average (ARIMA) (Dolara, leva and Manzolini,2015) and the numerical weather prediction (NWP) methods (Wu & Wu, 2020; Burnham, Anderson and Huyvaert, 2011). ARIMA is a prediction method based on Statistical model, which is a time series-based forecasting method that models and forecasts time series data (Boland, David and Lauret, 2016). The ARIMA model offers the advantages of simplicity, ease of use, and applicability to various types of time series data. Nonetheless, it has drawbacks including sensitivity to outliers and the requirement of extensive historical data. Statistical methods, despite being simpler than physical methods, can yield more accurate forecasting results due to their utilization of historical PV data and optimized model parameters. However, they still exhibit some notable shortcomings such as the need for a consistent correlation between inputs and outputs, resulting in lower accuracy on rainy and cloudy days. A prerequisite is the handling of large volumes of historical data, which can pose challenges during the data acquisition and training process. The effectiveness of statistical methods relies heavily on the quality and granularity of the PV data. Moreover, they are unable to extract intricate features associated with the inputs and PV power. Another method is Numerical Weather Prediction (NWP), which utilizes a physical model to forecast future weather conditions and, subsequently, estimate the future power output (Sun and Zhang, 2017). This is achieved through numerical simulations of atmospheric circulation and radiative transfer processes. For this method, historical time-series data are not required but detailed geological state of the plant, accurate meteorological weather data, and PV battery specifications are used. Although the forecasting accuracy highly depends on the NWP results, NWP performance may be reduced in some cases. In addition, the correlation between the model and PV panel is obtained with errors, the whole model is complex, and the computational cost is high. The advantage of the NWP model is that it can take more factors into account, which can improve the prediction accuracy. However, it also

[*] Corresponding author
Email: day_qs@163.com (Q. Dai)

https://doi.org/10.14710/ijred.2023.57902

has some disadvantages, such as high computational complexity and the need for a large number of computational resources.

Compared to traditional prediction methods, currently popular methods such as machine learning methods and deep learning methods have become widely used in various fields (Smith, 2018; Bochie *et al.*, 2021). These methods employ iterative training algorithms, allowing for the achievement of superior results using fewer computational resources (Rizk & Awad, 2019). Additionally, they possess the capability to process intricate data types and models, leading to greater precision in predicting outcomes (Petropoulos *et al.*, 2022). The Artificial Neural Network (ANN) is a type of machine learning method that consists of interconnected neurons or units. Each neuron receives inputs and outputs signals, generating an output by calculating a weighted sum of these inputs (Ragmani *et al.*, 2020). Given historical data containing internal and external variables, artificial neural networks (ANN) can be trained through supervised learning to predict future irradiance (Voyant *et al.*, 2014). SVM was used to achieve high accuracy by obtaining cloud covering position and cloudiness index (Peng *et al.*, 2015). When forecast lead time exceeds 1 hour, the machine learning techniques perform superior to their counterparts, and the disparity is more pronounced under cloudy conditions (Lauret *et al.*, 2015). However, machine learning methods are very sensitive to the temporal distance and the fineness of received input (Graditi, Ferlito and Adinolfi, 2016). Using ANNs to model linear trends may produce mixed results, therefore, blindly applying ANNs to any type of data is not advisable (Taskaya-Temizel & Casey, 2005).

Sun *et al.* first developed a convolutional neural networks (CNN) algorithm in 2018 that predicts the concurrent power output of solar panels solely from given sky images (Nie & Zamzam, 2021). Based on this model, Sun *et al.* proposed an architecture that predicts the photovoltaic output 15 minutes later by obtaining sky images from the past 15 minutes and the photovoltaic panel output from the same time period (Zhang, Wang and Liu, 2020), achieving a prediction rate of 16%. Zhang *et al.* compared three different deep learning architectures (Zhang *et al.*, 2018), including Multilayer Perceptron (MLP), CNN, and Long Short-Term Memory (LSTM) for predicting future photovoltaic power output one minute in advance. They found that the LSTM model had an advantage in capturing the temporal dependencies in the data, achieving a prediction rate of 21%.

In reference (Ajith & Martínez-Ramón, 2021), a combination of depth-wise separable convolution and LSTM was used to predict multi-step solar irradiance from multi-channel images (Wang *et al.*, 2019). However, this approach required more data and did not fully utilize historical data. In reference (Qu, Qian and Pei, 2021), ALSTM (AM-based LSTM) was used to extract long and short-term memory for hourly forecasting, achieving higher accuracy. However, since the extraction of spatiotemporal features was done separately, the extracted features may be distorted during the sequence processing, weakening the intrinsic correlation within local features. Reference (Trigo-González *et al.*, 2023) analyzed ultra-short-term photovoltaic power generation using data measured in Kyoto, Japan. Three convolutional neural network models were compared: Multilayer Perceptron. Convolutional Neural Network, and Long Short-Term Memory, with historical photovoltaic power values and sky images as inputs (Limouni & Yaagoubi, 2022). The authors considered the LSTM-based model to be superior to all other methods at that time (Tyass & Khalili, 2023; Nhat & Huu, 2023). However, the use of a single nonlinear regression can easily fall into local optima and has certain limitations.

The ground-based cloud prediction method analyses cloud information collected by ground-based all-sky imagers in order to assess the effect of clouds on solar radiation and to predict PV power (Sun *et al.*, 2014). This method is conducive to improving the accuracy of PV power prediction, especially in short- and ultrashort-term prediction, and has obvious advantages (Jaouhari, Zaz and Masmoudi, 2015; Lu, Wang and Li, 2021). However, the ground-based cloud mapping prediction method has limitations in temporal information utilization and is weak for migration between PV systems (Wei *et al.*, 2021). In the area of sequential data modelling, the transformer model with self-attention mechanism (Ma, 2022) is currently one of the most used neural network architectures in the field of Natural Language Processing (NLP), and was initially proposed for machine translation applications (Bi, Zhu and Meng, 2021). The revolution that the transformer architecture promoted in the NLP field motivated the development of new approaches based on the transformers in other areas such as computer vision and time series. For instance, in the computer vision area, (Liu, 2023) developed a Multiscale Vision Transformers (MViT) for video and image recognition, by connecting the seminal idea of multiscale feature hierarchies with transformer models. It efficiently captures information from the global context and is able to process all positions in the input sequence simultaneously (Xiao, Zhang and Ni, 2022). As a result, it can reduce computational time complexity and increase the speed of model training and inference (Fukushima & Ishikawa, 2022). However, there are a large number of parameters in the Transformer model, which requires a large amount of computational resources and storage space for training and inference (Nascimento *et al.*,2023; Fan H *et al.*,2021). In addition, VIT refers to Vision Transformer, a visual recognition model based on the Transformer architecture. It does this by splitting an image into a series of small squares (called patches) and then using the Transformer to process these patches, ultimately outputting a classification result for the entire image (Xiong *et al.*, 2020). Compared with traditional Convolutional Neural Networks (CNNs), the VIT model does not use convolutional and pooling layers, instead it is based entirely on the Transformer architecture for image classification. The VIT model automatically learns the relationship between each patch and is able to capture information inside the image over long distances, which makes the VIT a powerful performer in computer vision tasks. Although the VIT reduces the computational cost by encoding image data into vector representations, the VIT relies heavily on large-scale datasets for model training, which requires significant computational resources.

In this paper, a new deep neural network structure based on vision transformer (VIT) is proposed to combine sky images and Tokens-To-Token (T2T) for the above problems, which can gradually segment image tokenization into tokens with an efficient backbone. The T2T module is used to model the local structural information of an image while reducing the token length, and an efficient T2T-ViT backbone network is used to extract the global attentional relationships of the tokens from the T2T module.

2. Methodology

2.1 Data collection

This paper collected data from the Renewable Energy Climatology of NREL (National Renewable Energy Laboratory) in Golden, Colorado, USA. The dataset provides a

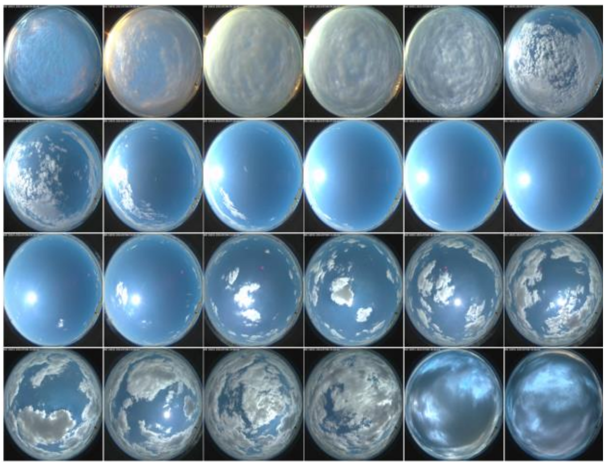

Fig 1. Cloud map data

comprehensive range of meteorological data, including wind speed and temperature. The measurement of cloud coverage was conducted by the NREL using a Solar Light 501A UVA radiometer, which recorded readings at 10-minute intervals. The radiometer measures wavelengths ranging from 315 nm in W/m², while maintaining a temperature of 25℃ during data collection. Additionally, an EKO All Sky Imager (ASI-16) was used to capture all sky images and compute cloud cover. This includes: a snapshot of the current sky conditions, updated every 60 seconds; total sky images and "cloud analyzed" images with corresponding cloud cover values every 10 minutes; percentages of thick and thin cloud cover every 10 minutes. Furthermore, the NREL utilized an EKO MS-300LR Sky Scanner to map luminance and irradiance at 15-minute intervals. The solar position or intensity was calculated based on the time and location using the Solar Position Calculators provided by the Measurement and Instrumentation Data Center (MIDC).Figure 1 provides examples of the data and illustrates the dynamic nature of the cloud diagrams.

2.2 Data preprocessing

In this study, data preprocessing is considered a critical phase for photovoltaic power forecasting, primarily aimed at ensuring the quality and continuity of the data required for the model. This paper employed cloud-based image transformation and linear interpolation methods. Initially, cloud image data was obtained in graphical format and subsequently subjected to enhancement, denoising, and computational filtering, enhancing the quality and consistency of the images. Subsequently, visual techniques were employed to extract relevant features from these cloud images, including cloud coverage percentage, cloud types, and cloud movement direction, all of which served as inputs to the model. Furthermore, by temporally aligning the extracted cloud image features with solar power generation data, this paper established a connection between meteorological information and electricity data. Additionally, to address missing data in solar radiation data, this paper adopted a linear interpolation method, estimating missing values by utilizing known data from neighboring time points. The significance of carefully managing the steps in handling missing data is emphasized in this paper to effectively mitigate noise and missing values in the dataset, consequently improving the accuracy and reliability of the T2T-vit model.

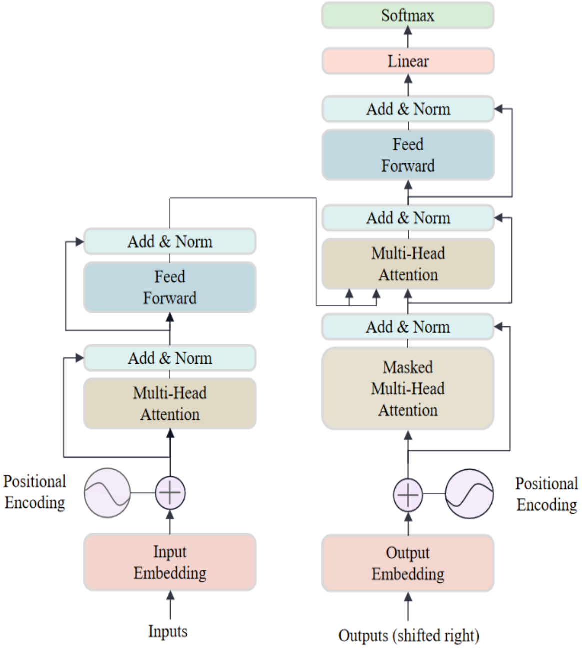

Fig 2. Transformer structure

2.3 Transformer

Transformer is a model that relies on a self-attention mechanism to map the global dependencies between inputs and outputs. Figure 2 illustrates the general Transformer structure. This model usually consists of a multi-head self-attention layer (MSA) and an MLP block, with LNs applied before each layer of self-attention, MLP block, and residual connections. Among them, the attention mechanism is the most central part . The attention mechanism is used as an alternative to the traditional Recurrent Neural Network (RNN) structure to capture the dependencies between different locations in the input sequence. The attention mechanism achieves better modelling capabilities and parallel computation by introducing a multi-head attention mechanism in the encoder and decoder. In the encoder, the attention mechanism helps the model to focus on a specific input word and weight it according to other words. In this way, each word gets a weighted representation of the surrounding context, enabling the model to better understand long-distance dependencies in a sentence. In the decoder, by introducing a self-attention mechanism, the model can associate generated parts of the target sequence with different parts of the input sequence. This enables the model to dynamically adjust the probability distribution of the output words according to the context and improve the generation accuracy.Recently the Transformer model has been applied to various vision tasks, image classification, target detection, image enhancement, image generation, video processing and so on. Among them vision Transformer (VIT) is the first Transformer model that can be directly applied to image classification by segmenting each image into 14×14 or 16×16 blocks (also known as tokens) of fixed length. Then after modelling with Transformer, the VIT applies the Transformer layer to model the global relationships between these tokens for classification. Although the VIT shows that the Transformer architecture is promising for visual tasks, it does not perform as well as similarly sized CNN models when training from scratch starting on a medium-sized dataset. Also the VIT relies heavily on large-scale datasets such as ImageNet-

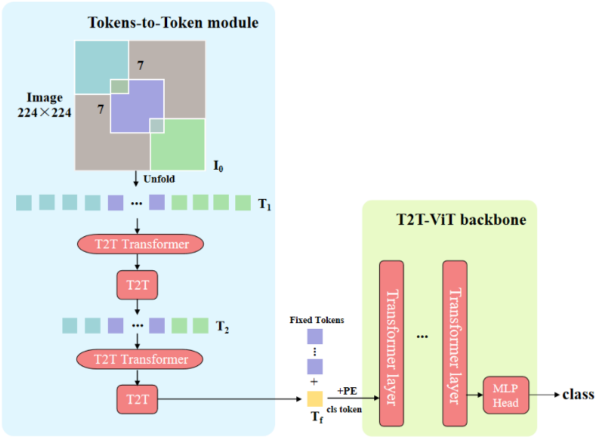

Fig 3. T2T-ViT transformer

21k and JFT-300M for model training, which requires significant computational resources.

The multi-head attention and scaled dot-product attention respectively use tree vectors: Queryvector (Q), a Key vector (K), and a Value vector (V). Transformersuse a "Scaled Dot-Product Attention" to obtain the context vector andcalculate the attention as:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V \qquad (1)$$

where Q $=W^Qx$, K = W^Kx, V = W^Vx on input x = $\{x_1, \ldots, x_n\}$. W^Q , W^K and W^V are weight matrices to generate Q, K and V via linear transformations on x. The head denotes the attention output result of each head, i.e., the weighted output result obtained after the calculation of the attention mechanism. Multi-head Attention is a module for attention mechanisms which allows the model to jointly attend to information from different representation subspaces in parallel, that is:

$$\text{MultiHead}(q, k, v) = Concat(head_1, head_2, \ldots, head_n).W^O \qquad (2)$$

where:

$$\text{head}_i = Attention(q.W_i^Q, k.W_i^K, v.W_i^V) \qquad (3)$$

2.4 Tokens-to-token vision transformer

In order to overcome the limitations of simple tokenization and inefficient backbone of the VIT, the T2T-ViT is proposed, which can gradually segment image tokenization into tokens with efficient backbone. The T2T module is used to model the local structural information of the image while reducing the token length, and efficient T2T-ViT backbone network is used to extract the global attentional relations of the tokens from the T2T module.

2.4.1 Tokens-to-token module

Instead of simple tokenization in VIT, the T2T-ViT employs an incremental tokenization module to aggregate neighbouring tokens into a single token (called Tokens-to-Token), which models the local structural information of the surrounding tokens and reduces the iterative token length. Specifically, in each T2T step, the token output from the Transformer layer is reconstructed as an image, then split into overlapping tokens, and finally aggregated into surrounding tokens via flatten. Thus, the local structures from the surrounding patches are embedded into the tokens of the next Transformer layer, and the local structures are aggregated into tokens by iterating T2T, and the length of the tokens can be reduced by the aggregation process.

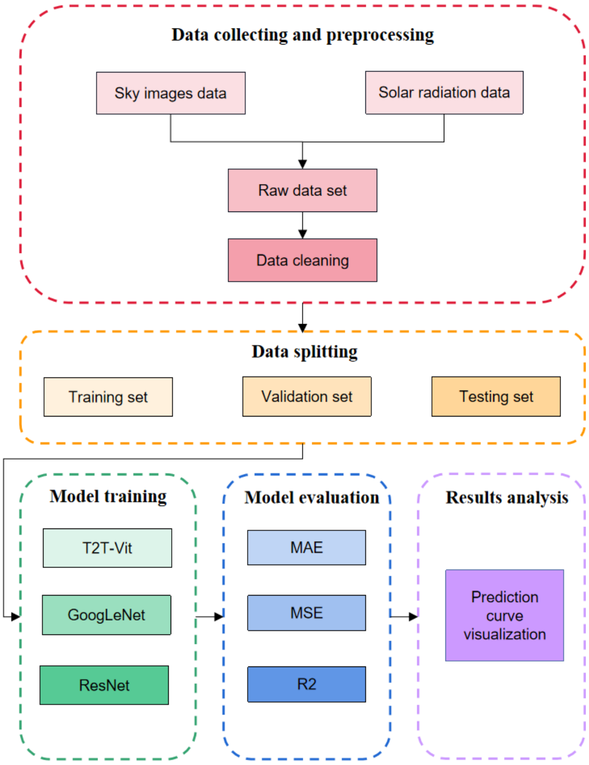

Fig 4. Predictive framework

2.4.2 T2T-Vit backbone

Since many channels in the trunk of vanilla the VIT are ineffective, there is a need to find an efficient Transformer trunk to reduce redundancy and improve feature richness. Therefore, different architectural designs for the VIT are explored to improve the efficiency of the backbone and enhance the richness of the learnt features, drawing on some of the designs of CNNs. By investigating five architectural designs from CNN to VIT: dense connectivity as in DenseNet; deep-narrow and shallow-wide structures as in Wide ResNet; channel attention as in Squeeze-Excitement (SE) network; more splitting heads in the multi-head attention layer as in ResNeXt; and Ghost operation as in Ghost Net, it is found that the use of deep-narrow structure that simply reduces the channel dimensions can reduce the channel redundancy in the VIT, and increasing layer depth improves feature richness in the VIT, with a reduction in model size and MACs, but with improved performance. In addition the channel attention of the SE block also improves the VIT, but not as effectively as the deep and narrow structure. Based on these findings, a deep-narrow structure is used for the T2T-VIT backbone. Specifically, it has a smaller number of channels and hidden dimensions d, but more layers b. For the last layer of fixed-length Token from the T2T block, a class Token is attached to it and then a sinusoidal positional embedding (PE) is added to perform the same classification as the VIT.

2.5 Experimental design and evaluation criteria

In this study, the collected Sky images data and Solar radiation data are used as raw data for data preprocessing and then divided into training set, testing set and validation set for model training. The effectiveness of the proposed prediction model is assessed by comparing its performance with well-established and mature deep neural network (DNN) structures like ResNet and GoogleNet. The accuracy of the model predictions was assessed using three error evaluation criteria: mean absolute error (MAE), mean square error (MSE), and coefficient of determination (R2). Subsequently, we visualize the prediction curves. The smaller the difference between the predicted and measured values, the better the model's prediction results.

In order to evaluate the forecasting model performance and improve the training process, this paper cites several error evaluation criteria. The MSE measures the predictive accuracy of a model by calculating the difference between the predicted and true values. Specifically, the MSE is the average value obtained by adding the squared differences of each data point. In addition, the mean square error (MSE) gradient decreases as the error decreases, which promotes function convergence. The function can rapidly achieve its minimum value when the learning rate remains constant. However, when outliers exist within the sample, MSE assigns a higher weight to these outliers, making the metric highly sensitive and significantly influenced by their presence. The MAE measures the average absolute error between the predicted value and the true value. The smaller the value of MAE, the better the model performance. This is because a smaller MAE value means that the model predicts less error. The input value exhibits a stable gradient regardless of its magnitude, ensuring that it does not result in the gradient explosion issue. Therefore, it possesses a fairly robust solution. While the MAE curve maintains continuity, it lacks differentiability at x = 0. Additionally, the MAE gradient remains uniform in the majority of cases, leading to significant gradients for even minor loss values. This characteristic hampers function convergence and impedes model learning. The R-square (R2) represents the portion of the variation in the dependent variable that can be explained by the model. It usually takes a value in the range of 0 to 1. The closer it is to 1, the better the model's ability to explain the data. And the predicted values built by the model fit the actual observations better. It is important to note that R2 does not tell us whether the model is statistically significant or not, but is merely an indicator that explains the goodness of fit of the model.

$$\mathrm{MSE} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \quad (4)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (5)$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad (6)$$

where i indicates the index of the sample data, ranging from 1 to n. n indicatins the total number of samples in the data set. y_i denotes the true value (or observed value) of the i-th sample. $\hat{y}_i$ denotes the i-th sample of the predicted value, i.e., the model's estimate of the true value.

3. Experiments and analysis of results

3.1 Mode structure and hyper-reference configuration

The experiments were conducted using an NVIDIA 4090Ti GPU with the PyTorch deep learning framework on a server. The initial learning rate was set to 0.0001 and gradually decreased to 0.00006 using the learning rate decay method. The initial learning rate is based on extensive experimentation and adjustments, and the learning rate attenuation method is utilized to facilitate model convergence towards the optimal solution

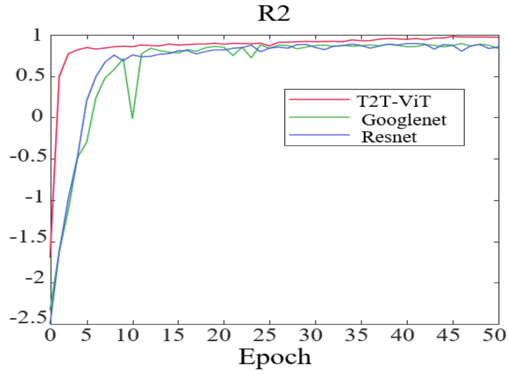

Fig 5. R2 of training

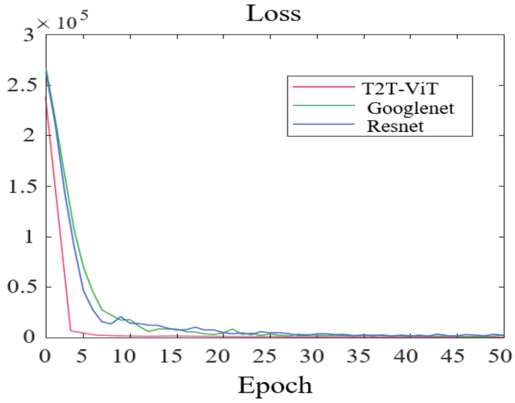

Fig 6. Loss of training

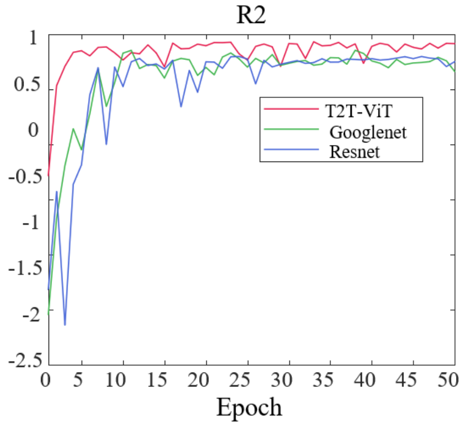

Fig 7. R2 of validation

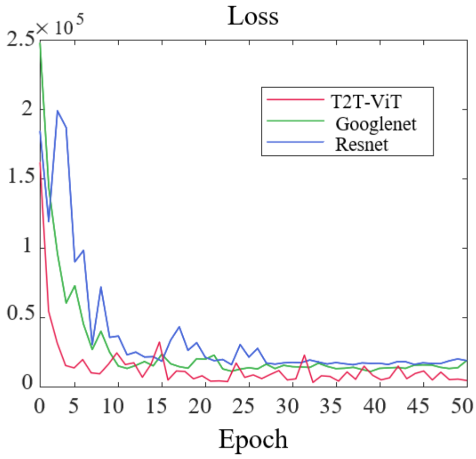

Fig 8. Loss of validation

Table 1
Predicted performance

Model	Error evaluation		
	Loss	MAE	R2
Googlenet	24.676	124.168	0.784
Resnet	42.041	157.045	0.633
T2T-ViT	**11.632**	**74.594**	**0.915**

during training. To mitigate gradient saturation, we employed the Adam optimizer instead of the traditional SGD optimizer. The Adam optimizer balances the gradient direction and learning rate step size by integrating the properties of AdaGrad and R MSProp. Additionally, we parameterized the comparison model to enable a comparative evaluation against the T2T-Vit model.

3.2 Comparison of experimental results

3.2.1 training and validation progress comparison

The Fig 6 and Fig 8 illustrates the dynamic loss of the three models during training and validation. As the training period increases, the loss decreases, indicating an improvement in the prediction accuracy of the models. Furthermore, the figure illustrates that the training loss of all three models exhibits a general declining trend, albeit with fluctuations. This phenomenon arises because the direction of gradient descent du ring each iteration of neural network training does not necessarily lead to the optimal solution as a whole. Consequently, the loss may not always decrease compared to the previous iteration. But, both on the training dataset and the validation dataset, the T2T-Vit model's loss converges faster than the Goodlenet model and the Resnet model, and the final loss values are smaller than the other models. Meanwhile, the comparison of the trends of the R^2 curves in Fig 5 and Fig 7. also shows the same results. This intuitively proves the advantages of our proposed model.

The Table 1 gives their respective error values, from which it can be seen that the Loss value of T2T-Vit is 11.632 is much s maller than 42.041 and 24.676, which proves that the error of T2T-Vit is smaller than the other two models. In addition, comparing the MAE value and R2 value of the three models can also prove the accuracy of T2T-Vit prediction.

3.2.2 Comparison of Forecast Results Presentation

To observe the prediction performance intuitively, we conducted tests on four consecutive days of data. The corresponding prediction performances are illustrated in Figures 9, 10, and 11, while the error of prediction is depicted in Figure 12. Specifically, the cloud image data selected for analysis spans from July 6th, 2021, to July 9th, 2021, encompassing diverse meteorological conditions. It is evident

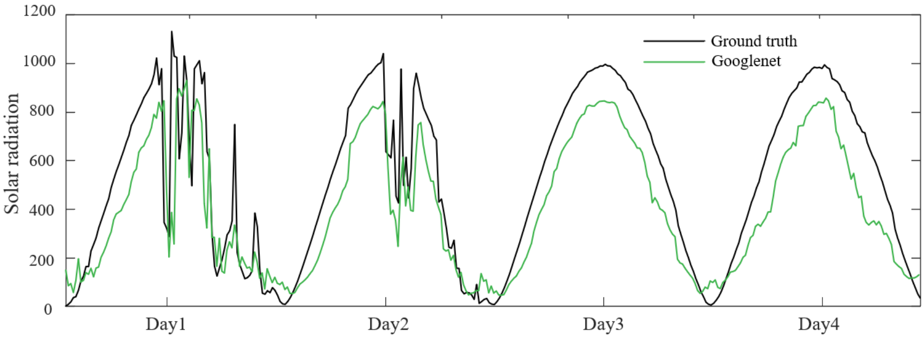

Fig 9. Predicted results of Googlenet

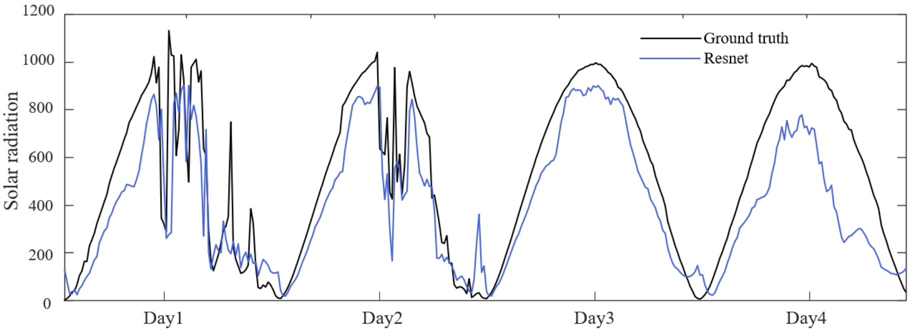

Fig 10. Predicted results of Resnet

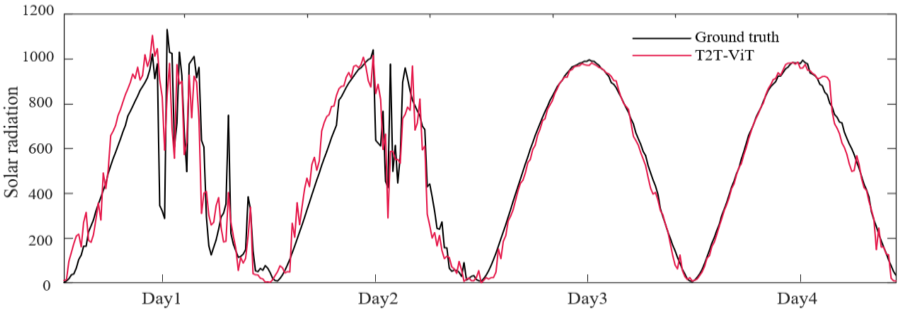

Fig 11. Predicted results of T2T-ViT

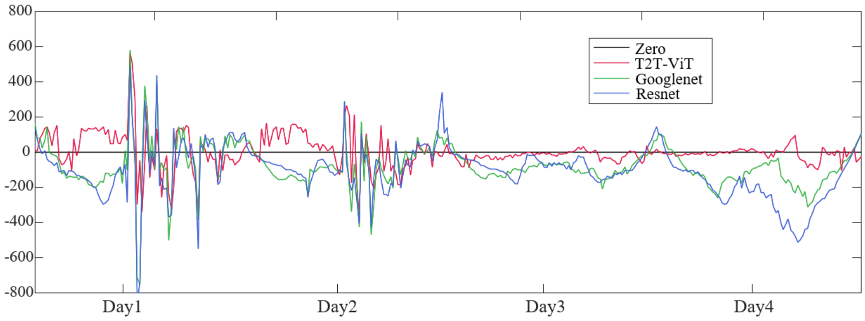

Fig 12. Comparison of model prediction error

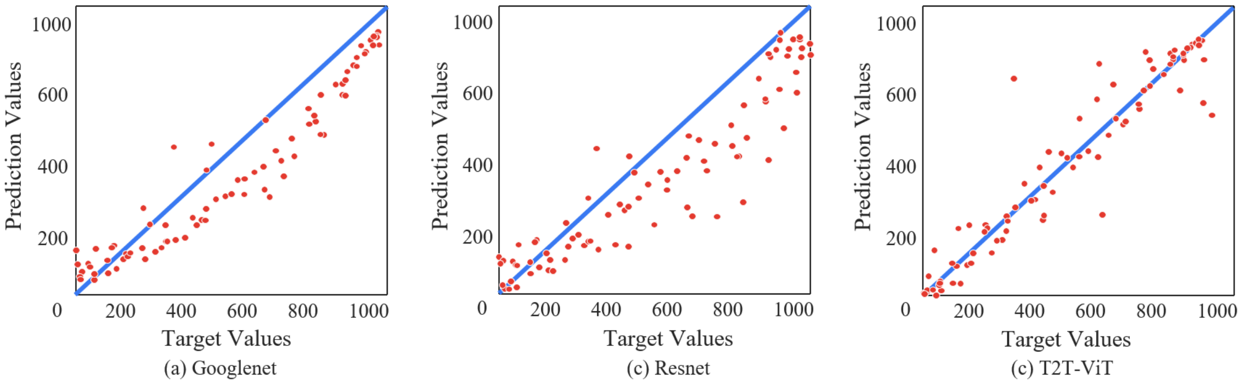

Fig 13. Comparison of model linear regression

that the proposed T2T-Vit model achieves higher prediction accuracy. However, it is worth noting that all models exhibit relatively more accurate predictions under low irradiance conditions. This phenomenon arises due to the inherent limitations of ground-based cloud images in accurately sampling cloud layers during periods of reduced solar light. Nevertheless, considering the limited power generation capacity and the resulting negligible fluctuations of photovoltaic equipment, this limitation can be deemed inconsequential.

Figure 13 shows the regression plots for three models, demonstrating the degree of fit between the data and the regression line. It is evident that, in comparison to the other two models, only a small number of data points in the T2T-Vit model exhibit minor deviations from the regression line, while the remaining data points are evenly dispersed on both sides of the regression line. Further demonstrating the superior prediction effect of the proposed model.

3.2.3 Advantages and disadvantages analysis of Forecast Results Presentation

The utilization of multiple Inception modules in GoogLeNet introduces elevated computational complexity, which in turn hinders efficient model computation. Furthermore, it is susceptible to issues such as gradient vanishing or explosion during training. Another drawback arises from the adoption of multi-scale convolutional kernels and pooling operations in GoogLeNet, as this may lead to redundant extraction of certain features across various convolutional levels, while disregarding other features.

ResNet employs a deeper network structure that involves multiple residual blocks, thereby yielding high computational complexity. Additionally, the method in which the residual blocks are connected in ResNet results in a greater number of model parameters. Consequently, increased memory is required to store these parameters, while more computational resources are necessary for training and inference. Furthermore, the existence of shortcut connections in residual blocks may lead to limited utilisation of certain feature levels, reducing the network's capacity to effectively reuse its underlying features.

In contrast, the T2T-ViT approach employed in this study proficiently captures the global spatial relationship and contextual information within an input image via the self-attention mechanism. This confers a notable advantage to T2T-ViT when addressing visually intensive tasks involving long-range dependencies and global awareness, such as target detection and image segmentation. Additionally, T2T-ViT incorporates a multi-scale input and feature fusion strategy, enabling effective handling of input images of varying sizes. Moreover, the Transformer structure of T2T-ViT appears relatively simple, with each attention head capable of providing interpretable feature representations. Furthermore, this study integrates sky images into the framework. By combining sky images with other image data, a more comprehensive understanding and interpretation of elements such as objects, actions, and backgrounds within the images can be achieved. Simultaneously, weather elements like clouds, light, and color present in the sky image contribute additional features to this model. These features enhance the image representation by encompassing more visual cues, ultimately bolstering prediction performance.

4.Conclusion

This paper presents a novel approach for predicting photovoltaic power using sky images and Tokens-To-Token Vision Transformer (T2T-ViT). The approach employs an incremental tokenization module to combine adjacent image patches into tokens, which capture the local structural information of the clouds. Subsequently, an efficient T2T-ViT backbone network is utilized to extract the global attentional relationships among the tokens for power prediction. To evaluate the effectiveness of the proposed method, a dataset from the National Renewable Energy Laboratory in Colorado, USA is employed for comparison with various deep learning architectures, including ResNet and GoogleNet. The experimental results demonstrate that the proposed method achieves higher prediction accuracy and lower error compared to existing methods, particularly in short- and ultra-short-term predictions.

Author Contributions: Qiangsheng Dai.: Conceptualization, methodology, formal analysis, writing—original draft, Xuesong Huo.; supervision, resources, formal analysis, Dawei Su.; writing—review and editing, project administration, validation, Zhiwei Cui.; writing—review and editing, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by science and technology project of State Grid JiangSu Electric Power Co., Ltd. (The research of distribution network distributed photovoltaic high-precision prediction and local consumption capacity, Grant No. is J2022043).

Conflicts of Interest: The authors declare no conflict of interest.

References

Agoua, X. G., Girard, R., & Kariniotakis, G. (2018). Short-term spatio-temporal forecasting of photovoltaic power production. *IEEE Trans. Sustain. Energy,* 9(2), 538–546. https://doi.org/10.1109/ICPES.2017.8387332

Ajith, M. & Martínez-Ramón, M. (2021). Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data, *Applied Energy*, 294. https://doi.org/10.1016/j.apenergy.2021.117014.

Bi, J., Zhu Z., & Meng, Q. (2021). Transformer in Computer Vision. *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI),* Fuzhou, China, pp. 178-188 https://doi.org/10.1109/CEI52496.2021.9574462.

Biswas, A. K., Ahmed, S. I., Bankefa, T., Ranganathan, P., and Salehfar, H. (2021). Performance Analysis of Short and Mid-Term Wind Power Prediction using ARIMA and Hybrid Models. *IEEE Power and Energy Conference at Illinois (PECI)*, Urbana, IL, USA, 2021. https://doi.org/10.1109/PECI51586.2021.9435209

Bochie, K., Gilbert, M. S., Gantert, L., Barbosa, M. S. M., Medeiros, D. S. B., & Campista, M. E. M. (2021). A survey on deep learning for challenged networks: Applications and trends, *Journal of Network and Computer Applications*, 194, 2021. https://doi.org/10.1016/j.jnca.2021.103213.

Boland, J., David M, & Lauret P. (2016). Short term solar radiation forecasting: island versus continental sites. *Energy* .113:186e92. https://doi.org/10.1016/j.energy.2016.06.139

Burnham, K., Anderson D., & Huyvaert KK. (2011). Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol.*65:23e35. https://doi.org/10.1007/s00265-010-1029-6

Dolara, A., Leva, S., & Manzolini, G. (2015). Comparison of different physical models for pv power output prediction. *Sol Energy.* 119:83e99. https://doi.org/10.1016/j.solener.2015.06.017

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscalevision transformers. *In: Proceedings of the IEEE/CVF international conference oncomputer* vision. https://www.sciencedirect.com/science/refhub/S0360-5442(23)01072-1/sb14

Fukushima, D. & Ishikawa, T. (2022). Experiments And Discussions On Vision Transformer (ViT) Parameters For Object Tracking, *2022 Nicograph International (NicoInt),* Tokyo, Japan, pp. 64-68 https://doi.org/10.1109/NicoInt55861.2022.00020.

Gao, Y., Shi, S., Sun, Z. , & Ling, C.. (2022). The combination of transformer and CNN in computer vision," 2022 *IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT),* Dali, China, pp. 321-325 https://doi.org/10.1109/ICCASIT55263.2022.9987025

Graditi, G., Ferlito, S., & Adinolfi, G. (2016). Comparison of Photovoltaic plant power production prediction methods using a large measured dataset, *Renew Energy*, 90, 513-519. https://doi.org/10.1016/j.renene.2016.01.027

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D. (2022). A Survey on Vision Transformer. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87-110. https://doi.org/10.1109/TPAMI.2022.3152247.

Jaouhari, Z. El, Zaz, Y., & Masmoudi, L. (2015). Cloud tracking from whole-sky ground-based images," *2015 3rd International Renewable and Sustainable Energy Conference (IRSEC),* Marrakech, Morocco, pp. 1-5 https://doi.org/10.1109/IRSEC.2015.7455105.

Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P. (2015). A benchmarking of machine learning techniques for solar radiation forecasting in an insular context, *Sol Energy,* 112, 446-457. https://doi.org/10.1016/j.solener.2014.12.014

Liu, J., Zang, H., Cheng, L., Ding, T., Wei, Z., & Sun, G., (2023). A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting, *Applied Energy,* 342, 121160, https://doi.org/10.1016/j.apenergy.2023.121160.

Liu, S. (2023). Application and Analysis of Convolutional Neural Networks and Vision Transformer Models in Fruit Recognition, *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA),* Shenyang, China, pp. 1530-1533. https://doi.org/10.1109/ICPECA56706.2023.10076218.

Liu, W., Ren, C., & Xu, Y. (2021). PV generation forecasting with missing input data. A super-resolution perception approach. *IEEE Trans. Sustain. Energy,* 12(2), 1493–1496. https://doi.org/10.1109/TSTE.2020.3029731

Limouni, T., Yaagoubi, R. (2022). Univariate and Multivariate LSTM Models for One Step and Multistep PV Power Forecasting. *International Journal of Renewable Energy Development*, 11(3), 815-828. https://doi.org/10.14710/ijred.2022.43953

Lu, Z., Wang, K., & Li, X. (2021). A Preprocessing and Feature Extraction Method of Ground-based Cloud Images for Photovoltaic Power Prediction. *2021 40th Chinese Control Conference (CCC),* Shanghai, China, pp. 6823-6828 https://doi.org/10.23919/CCC52363.2021.9549239.

Ma, D., (2022). Recent advances in deep learning based computer vision. *2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE),* Ottawa, ON, Canada, pp. 174-179 https://doi.org/10.1109/CIPAE55637.2022.00044.

Nascimento E. G. S., Talison A.C. de Melo, Davidson M. Moreira. (2023). A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy. *Energy.* https://doi.org/10.1016/j.energy.2023.127678.

Nhat, N. N. V., Huu, D. N. (2023). Evaluating the EEMD-LSTM model for short-term forecasting of industrial power load: A case study in Vietnam. *International Journal of Renewable Energy Development,* 12(5), 881-890https://doi.org/10.14710/ijred.2023.55078

Nie, Y., & Zamzam, Ahmed S. (2021). Adam Brandt,Resampling and data augmentation for short-term PV output prediction based on an imbalanced sky images dataset using convolutional neural networks, *Solar Energy*, 224, 341-354. https://doi.org/10.1016/j.solener.2021.05.095.

Peng, Z., Yu, D., Huang, D., Heiser, J., Yoo, S., &Kalb, P. (2015). 3D cloud detection and tracking system for solar forecast using multiple sky imagers, *Sol. Energy* 118, 496–519. https://doi.org/10.1016/j.solener.2015.05.037

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., & Barrow, D. K. (2022). Souhaib Ben Taieb Forecasting: theory and practice, *International Journal of Forecasting*, 38(3), 705-871. https://doi.org/10.1016/j.ijforecast.2021.11.001.

Qu, J., Qian, Z., & Pei, Y., (2021). Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern, *Energy,* 232. https://doi.org/10.1016/j.energy.2021.120996.

Ragmani, A., Elomri, A., Abghour, N., Moussaid, K., Rida, M., & Badidi, E. (2020). Adaptive fault-tolerant model for improving cloud computing performance using artificial neural network, *Procedia Computer Science,* 170. 929-934. https://doi.org/10.1016/j.procs.2020.03.106.

Rizk, Y., & Awad, M.(2019).On extreme learning machines in sequential and time series prediction. A non-iterative and approximate training algorithm for recurrent neural networks, *Neurocomputing,* 325. 1-19, https://doi.org/10.1016/j.neucom.2018.09.012.

Smith, A. (2018). Machine Learning – A Review. *Journal of Artificial Intelligence Research,* 45(3). 589-604. https://doi.org/10.1002/jls.21605

Sun, S., Ernst, J., Sapkota, A., Ritzhaupt-Kleissl, E., Wiles, J., Bamberger, J., Chen, T. (2014). Short term cloud coverage prediction using ground based all sky imager, *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm),* Venice, Italy, pp. 121-126. https://doi.org/10.1109/SmartGridComm.2014.7007633.

Sun, X., & Zhang, T. (2017). Solar Power Prediction in Smart Grid Based on NWP Data and an Improved Boosting Method. *2017 IEEE International Conference on Energy Internet (ICEI),* Beijing, China. pp. 89-94.https://doi.org/10.1109/ICEI.2017.23

Takeda, M. & Yanai, K., (2022). Continual Learning in Vision Transformer. *2022 IEEE International Conference on Image Processing(ICIP),* Bordeaux, France, pp. 616-620. https://doi.org/10.1109/ICIP46576.2022.9897851.

Taskaya-Temizel, T., Casey, M. C. (2005). A comparative study of autoregressive neural network hybrids *Neural Networks*, 18 (5-6) 781-789. https://doi.org/10.1016/j.neunet.2005.06.003

Trigo-González, M., Cortés-Carmona, M., Marzo, A., Alonso-Montesinos, T., Martínez-Durbán, M., López, G., Portillo, C., & Batlles, F. J. (2023). Photovoltaic power electricity generation nowcasting combining sky camera images and learning supervised algorithms in the Southern Spain, *Renewable Energy*, 206, 251-262 https://doi.org/10.1016/j.renene.2023.01.111.

Tyass, I. & Khalili, T. (2023). Wind Speed Prediction Based on Statistical and Deep Learning Models. *International Journal of Renewable Energy Development,* 12(2), 288-291 https://doi.org/10.14710/ijred.2023.48672

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., & Kaiser, Ł., & Polosukhin I. (2017). Attention is all you need. In: Advances in neural informationprocessing systems. pp. 5998–6008. https://refhub.elsevier.com/S0360-5442(23)01072-1/sb12

Voyant, C., Haurant, P., Muselli, M., Paoli, C., &Nivet, M.-L. (2014). Time series modeling and large scale global solar radiation forecasting from geostationary satellites data, *Sol Energy*, 102, 131-142. https://doi.org/10.1016/j.solener.2014.01.017

Wang, X., Kan, M., Shan, S., & Chen, X. (2019). Fully Learnable Group Convolution for Acceleration of Deep Neural Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Long Beach, CA, USA. pp. 9041-9050. https://doi.org/10.1109/CVPR.2019.00926

Wei, L., Zhu, T., Guo, Y., Ni C., and Zheng, Q. (2023). CloudpredNet: An Ultra-Short-Term Movement Prediction Model for Ground-Based Cloud Image. in *IEEE Access* https://doi.org/10.1109/ACCESS.2023.3310538.

Wu S. -H. & Wu, Y. -K. (2020). Probabilistic Wind Power Forecasts Considering Different NWP Models. *2020 International Symposium on Computer, Consumer and Control (IS3C),* Taichung City, Taiwan. pp. 428-431. https://doi.org/10.1109/IS3C50286.2020.00116

Xiao, Y., Zhang, Y., & Ni, P. (2022). Ensemble Long Short-Term Tracking with ConvNeXt and Transformer. *2022 7th International Conference on Image, Vision and Computing (ICIVC),* Xi'an, China, pp. 688-693 https://doi.org/10.1109/ICIVC55077.2022.9887117.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. (2020). On layer normalization in the transformer architecture. *In: International conference on machine learning. PMLR.* pp. 10524–33. http://refhub.elsevier.com/S0360-5442(23)01072-1/sb13

Zhang, H., Wang, L., & Liu, J. (2020). Comparative analysis of deep learning architectures for predicting future PV power output. *Solar Energy*, 456(3), 78-92. https://doi.org/10.1016/j.energy.2021.122733

Zhang, J., Verschae, R., Nobuhara, S., & Lalonde, J., (2018). Deep photovoltaic nowcasting, *Solar Energy*, 176. 267-276. https://doi.org/10.1016/j.solener.2018.10.024.

Zhang, L., Wilson, R., Sumner, M., & Wu, Y., (2023). Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach, *Renewable Energy*, 216, 118952, https://doi.org/10.1016/j.renene.2023.118952

Zhang, Y., Qian, W., Ye, Y., Li, Y., Tang, Y., Long, Y., & Duan, M. (2023). A novel non-intrusive load monitoring method based on ResNet-seq2seq networks for energy disaggregation of distributed energy resources integrated with residential houses, *Applied Energy*, 349, 121703. https://doi.org/10.1016/j.apenergy.2023.121703.