

Article

Transformer with Adaptive Sparse Self-Attention for Short-Term Photovoltaic Power Generation Forecasting

Xingfa Zi ^{1,*}, Feiyi Liu ¹, Mingyang Liu ¹ and Yang Wang ²¹ School of Physics, Electrical and Energy Engineering, Chuxiong Normal University, Chuxiong 675000, China; fylu@cxtc.edu.cn (F.L.); liumingyang@cxtc.edu.cn (M.L.)² School of Big Data and Basic Science, Shandong Institute of Petroleum and Chemical Technology, Dongying 257061, China; wangyang@sdipct.edu.cn

* Correspondence: zxf@cxtc.edu.cn

Abstract

Accurate short-term photovoltaic (PV) power generation forecasting is critical for the stable integration of renewable energy into the grid. This study proposes a Transformer model enhanced with an adaptive sparse self-attention (ASSA) mechanism for PV power forecasting. The ASSA framework employs a dual-branch attention structure that combines sparse and dense attention paths with adaptive weighting to effectively filter noise while preserving essential spatiotemporal features. This design addresses the critical issues of computational redundancy and noise amplification in standard self-attention by adaptively filtering irrelevant interactions while maintaining global dependencies in Transformer-based PV forecasting. In addition, a deep feedforward network and a feature refinement feedforward network (FRFN) adapted from the ASSA-Transformer are incorporated to further improve feature extraction. The proposed algorithms are evaluated using time-series data from the Desert Knowledge Australia Solar Centre (DKASC), with input features including temperature, relative humidity, and other environmental variables. Comprehensive experiments demonstrate that the ASSA models' accuracy in short-term PV power forecasting increases with longer forecast horizons. For 1 h ahead forecasts, it achieves an R^2 of 0.9115, outperforming all other models. Under challenging rainfall conditions, the model maintains a high prediction accuracy, with an R^2 of 0.7463, a mean absolute error of 0.4416, and a root mean square error of 0.6767, surpassing all compared models. The ASSA attention mechanism enhances the accuracy and stability in short-term PV power forecasting with minimal computational overhead, increasing the training time by only 1.2% compared to that for the standard Transformer.



Academic Editor: Fabio Corti

Received: 27 August 2025

Revised: 9 October 2025

Accepted: 10 October 2025

Published: 11 October 2025

Citation: Zi, X.; Liu, F.; Liu, M.; Wang, Y. Transformer with Adaptive Sparse Self-Attention for Short-Term Photovoltaic Power Generation Forecasting. *Electronics* **2025**, *14*, 3981. <https://doi.org/10.3390/electronics14203981>

Copyright: © 2025 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/>).

Keywords: deep learning; transformer; adaptive sparsity self-attention; renewable energy; photovoltaic energy; PV power generation forecasting; time-series forecasting

1. Introduction

The application potential of deep learning (DL) is increasingly evident in the photovoltaic (PV) industry. Its ability to process high-dimensional, multi-source, and heterogeneous data has made it indispensable for key tasks such as PV power prediction, fault detection, component identification, and operation and maintenance optimization [1,2]. Compared with the scalability and feature engineering limitations of traditional machine learning methods [3–6], neural network (NN) algorithms have significantly improved the accuracy of data processing and prediction [7–10]. In particular, deep learning techniques such

as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks can extract latent information from time-series data and capture the dependencies between historical and future observations, thereby enhancing forecasting performance [11–14]. Overall, NN-based approaches have opened new avenues for PV power forecasting and provided strong support for the intelligent and sustainable development of the PV industry [15–17].

The Transformer, a deep learning framework featuring an encoder–decoder architecture and a self-attention mechanism [18], has demonstrated considerable potential for PV power forecasting [19]. By modeling spatiotemporal dependencies, it effectively handles complex and highly variable weather patterns [20]. Subsequent research has introduced various Transformer variants to further enhance its applicability. For example, hybrid models combining one-dimensional convolutional neural networks (CNNs) with Transformer architectures enable multi-step forecasting across different temporal resolutions [21]. The PVTransNet model family, particularly the PVTransNet-EDR variant integrated with an LSTM, has achieved up to a 48.3% higher forecast accuracy compared with that of an LSTM alone [22]. More recently, Fourier graph neural networks (FourierGNNs) have extended this line of work by leveraging the Fourier transform to extract spatiotemporal dependencies from PV data represented as hypervariable graphs [23]. They integrate signal decomposition and sparse attention to address sudden weather changes [24], adopt shifted window-based architectures for transferable predictions across PV sites [25], and refine input features through specialized construction techniques [26,27]. Despite these advances, a common limitation remains. The standard self-attention mechanism computes interactions between all tokens in a sequence, which often introduces computational redundancy and amplifies noise in high-frequency time-series data. This reduces model efficiency and obscures critical features. Although existing methods improve forecasting performance, they primarily rely on auxiliary modules or feature engineering, leaving a fundamental gap: the redundancy inherent in self-attention and the absence of adaptive feature selection remain unresolved.

To address this limitation, we propose the adaptive sparse self-attention (ASSA) mechanism [28]. ASSA mitigates the inherent shortcomings of the standard attention mechanism by employing a two-branch structure: a sparse self-attention (SSA) branch that filters out low-relevance interactions and a dense self-attention (DSA) branch that preserves essential information. An adaptive weighting module dynamically integrates the outputs of the two branches, enabling the model to concentrate on the most informative time steps while suppressing noise and redundancy and maintaining the capacity to capture global dependencies. Consequently, this architectural enhancement improves model robustness and forecasting accuracy under noisy conditions while providing a more efficient research direction for Transformer-based PV forecasting.

This work proposes a Transformer-based framework that integrates the ASSA mechanism to improve short-term PV power forecasting. To systematically assess the effectiveness of ASSA, we design three architectural variants. The first variant substitutes the standard multi-head attention with an ASSA module, demonstrating its ability to adaptively suppress noise while retaining key spatiotemporal dependencies. The second variant, “ASSA Deep”, incorporates a bottleneck-structured deep feedforward network to strengthen the modeling of high-order nonlinear interactions. The third variant, “ASSA FRFN”, replaces the conventional position-based feedforward network with a feature refinement feedforward network that employs partial convolution and gating mechanisms, enabling dynamic fine-grained feature enhancement across both channel and temporal dimensions. We evaluate the proposed models against widely used deep learning baselines, including LSTM and the standard Transformer. Experiments on real-world PV power generation data show that the ASSA mechanism substantially improves forecasting accuracy under variable weather

conditions while incurring less computational overhead. This study establishes a robust and efficient framework for short-term PV power forecasting, thereby facilitating the stable integration of solar energy into smart grids.

This paper is organized as follows. Section 2 details the data preprocessing and the Transformer model with adaptive sparse self-attention (ASSA), including the standard module design and ASSA variants. Section 3 describes the dataset and processing methods. Section 4 presents performance comparisons of the proposed model for PV prediction at different time resolutions and under different weather conditions. Section 6 summarizes the main findings, practical implications, and future research directions.

2. Methodology

In this work, a Transformer-based architecture integrated with ASSA is employed to investigate its impact on PV power generation forecasting. This section presents the structure and parameter settings of the Transformer model, explains the mechanism and advantages of ASSA, and designs four algorithmic variants, including a baseline Transformer, its integration with ASSA, and its variants, to evaluate the performance improvements provided by ASSA.

2.1. Transformer, Adaptive Sparse Self-Attention, and Feature Refinement Feedforward Network

The Transformer is a well-known deep learning architecture based on self-attention, consisting of an encoder and a decoder, as illustrated in Figure 1. It models sequential dependencies through the integration of multi-head attention and a position-wise feed-forward network. After embedding and positional encoding, the input passes through stacked layers, each containing a self-attention module and a feedforward network with residual connections and layer normalization to stabilize training. The decoder extends this design by adding an extra attention module to process encoder outputs and uses masking to ensure autoregressive generation. The Transformer architecture shown in Figure 1 serves as the baseline in this study, with its mathematical formulation provided in Appendix A.1.

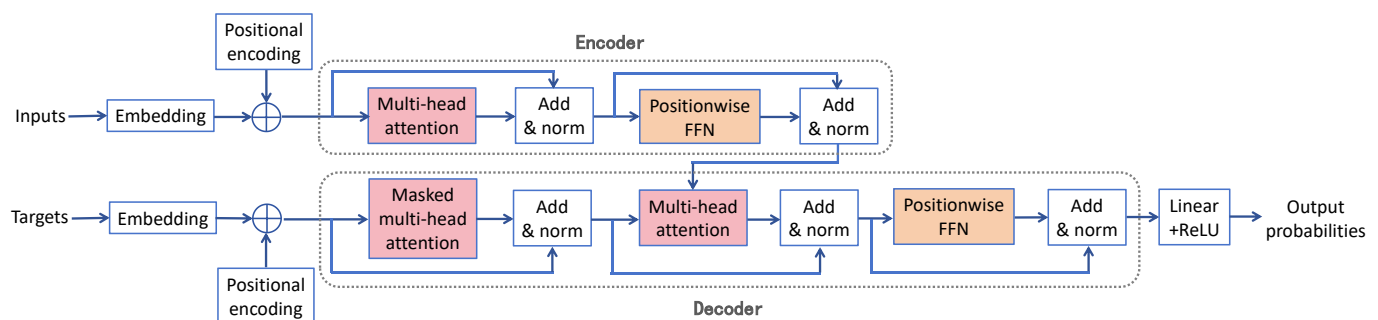


Figure 1. The architecture of the Transformer.

Adaptive sparse self-attention (ASSA) aims to mitigate noise interactions from irrelevant regions in the Transformer while preserving informative features [28]. Since the Transformer attends to all tokens in the feature map, it may involve redundant or irrelevant regions during the computation process [29,30]. To address this issue, the core innovation of ASSA lies in adaptively combining two complementary branches, which jointly suppress noise and redundancy while maintaining the integrity of critical information. A squared ReLU-based self-attention mechanism (SSA) is introduced to suppress features with low query–key matching scores, while a dense self-attention branch (DSA) applies a softmax operation in parallel to preserve essential information and compensate for potential oversparsity in the SSA, as shown in Figure 2. In essence, ASSA enables the model to adaptively control the sparsity of attention in task-specific scenarios, thereby enhancing robustness

and feature selectivity. Unlike conventional sparsification techniques that rely on selection or restriction, ASSA achieves sparsity through adaptive fusion, effectively balancing sparse and dense representations to yield task-adaptive soft sparsity. Consequently, ASSA serves not merely as a sparsification method but as an attention regulation mechanism. The core mathematical formulation of ASSA is provided in Appendix A.2.

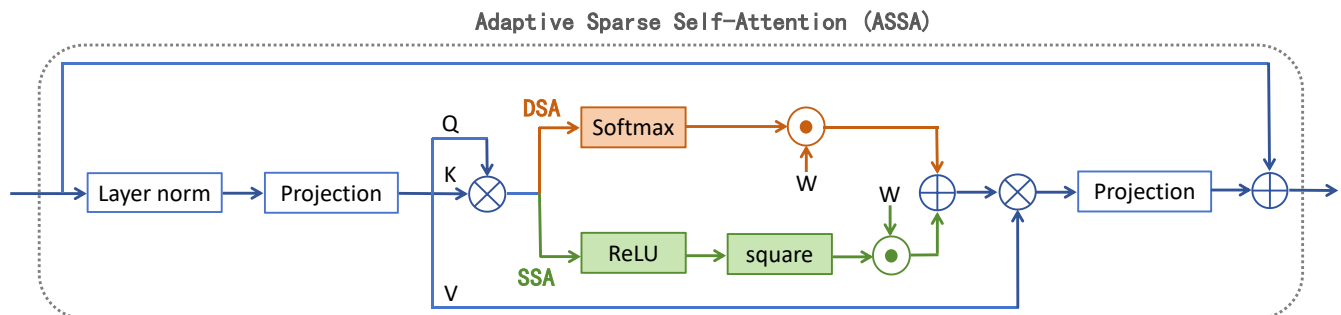


Figure 2. The architecture of ASSA.

Compared with the conventional multi-head self-attention mechanism in the Transformer, ASSA may exhibit several advantages in PV forecasting. ASSA significantly reduces the computational complexity by replacing the dense attention matrix with a sparse and adaptive structure, which can improve training and inference efficiency. ASSA enhances the modeling of long-term dependencies in PV time-series data by selectively focusing on the most informative time steps, thereby mitigating the noise interference. The sparsity and adaptivity of ASSA alleviate the risk of overfitting to short-term fluctuations, which could improve the generalization ability across different weather and illumination conditions. Therefore, this study investigates the application of the ASSA–Transformer architecture for PV forecasting. Specifically, we replace the standard self-attention mechanism in the Transformer with ASSA and employ a series of ASSA–Transformer variants for comparative analysis.

We also introduce the feature refinement feedforward network (FRFN) to optimize the feature representation by enhancing informative features and reducing redundancy along the channel dimension [28]. The main components of the FRFN include partial convolution (PartialConv), a gating mechanism, residual connections, and a nonlinear FNN, as illustrated in Figure 3. Specifically, the FRFN strengthens relevant feature elements through partial convolution and employs a gating mechanism to alleviate the computational burden caused by redundant information. Partial convolution applies lightweight operations to only a subset of input channels, thereby retaining key features while filtering out redundancy. This design substantially reduces the number of parameters and computational complexity, making it well suited to processing large-scale data [31]. The mathematical description of the FRFN part is given in Appendix A.3.

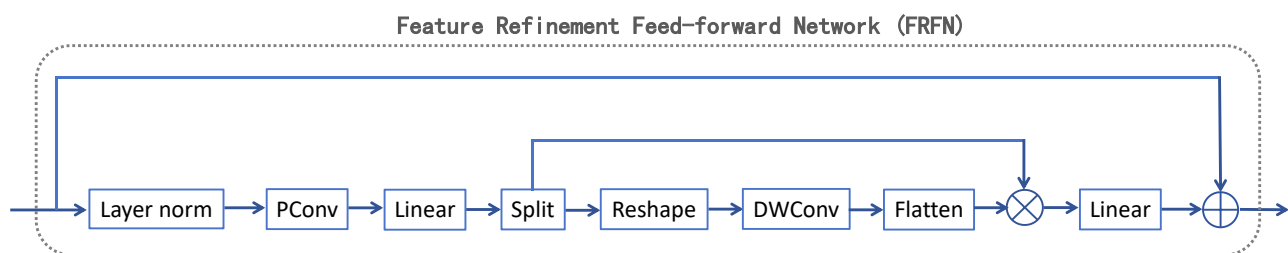


Figure 3. The architecture of the FRFN.

The FRFN is well suited to time-series forecasting due to its adaptive feature refinement capabilities. Dynamic feature selection is enabled as the model is directed by the partial convolution operation to focus on the most informative input feature combinations at each time step. Concurrently, a data-dependent temporal filter is constituted by the gating mechanism, whereby critical periods (e.g., weather transitions) are amplified and noisy intervals are suppressed. The refinement of the complex and noisy spatiotemporal patterns in PV power generation data is thereby effectively accomplished through this integration of sparse feature processing and adaptive gating, yielding more robust forecasts compared with standard FNNs. Consequently, the focus of this study is set on one-dimensional time series, where temporal dependencies are critical. To align with this focus, the positional FNN in Figure 1 is replaced by the FRFN, thereby forming a variant of the ASSA–Transformer. The complete ASSA FRFN algorithm from Ref. [28] is not directly adopted, and this modification is implemented to preserve the core FRFN functionality while placing greater emphasis on temporal feature processing.

2.2. Model Architecture

In this work, the Transformer-based ASSA algorithm and its variants are constructed for performance evaluation in PV power forecasting. The baseline is established using the standard Transformer described in Section 2.1. A simplified ASSA–Transformer is then obtained by replacing the multi-head attention with ASSA while the rest of the architecture is retained. Subsequently, the position-wise FFN is substituted with a deeper neural network for enhanced time-series feature extraction and is further replaced with the FRFN to assess the impact of a more complex architecture. The architectural designs of these ASSA–Transformer variants are detailed in the following sections, with a comprehensive presentation provided in this subsection.

The original input data is denoted as $X \in \mathbb{R}^{N \times D}$, where N is the number of time steps and D is the number of features at each step. The PV generation target is placed in the last column, while the remaining columns correspond to environmental variables. To ensure consistency in the feature magnitudes, improve feature learning efficiency, and enhance prediction accuracy, the data is normalized as

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (1)$$

where $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values in X , respectively. This linear mapping transforms the original data into the interval $[0, 1]$, mitigating the influence of feature scale differences on model training. Then, the normalized data is prepared for input into the model. Assuming an input window with a length W , a prediction step length of L , and a batch size of B , the algorithm input is defined as $X_{\text{input}} \in \mathbb{R}^{B \times W \times D}$.

To incorporate spatiotemporal information into the input sequence, the Transformer embeds and positionally encodes the data at each time step, providing the model with both semantic information and temporal position context. Each time step in the input sequence typically consists of multiple numerical values, which are projected into a specified embedding space. The input embedding can be expressed as

$$E = X_{\text{input}} W_e + b_e, \quad (2)$$

where $W_e \in \mathbb{R}^{D \times D_E}$ is the embedding weight matrix, and D_E denotes the embedding dimension. The bias vector $b_e \in \mathbb{R}^{D_E}$ encodes positional information using sine and cosine

functions, assigning a unique code to each position in the sequence. Specifically, for position p and dimension i ,

$$\begin{aligned} b_e(p, 2i) &= \sin\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right), \\ b_e(p, 2i + 1) &= \cos\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right). \end{aligned} \quad (3)$$

The positional encoding is directly added to the input embedding, injecting temporal information while preserving the original feature values. In this study, the embedding dimension is set to $D_E = 32$, resulting in an embedded input of size $E \in \mathbb{R}^{B \times W \times D_E}$. Next, linear projections are applied to the embedded data E to generate the inputs for the attention mechanism of the Transformer: the query matrix Q , the key matrix K , and the value matrix V . This can be expressed as

$$Q = EW^Q, \quad K = EW^K, \quad V = EW^V, \quad (4)$$

where W^Q , W^K , and $W^V \in \mathbb{R}^{D_E \times d_k}$ are learnable projection matrices, and d_k denotes the dimension of each attention subspace.

The first ASSA–Transformer architecture is constructed by replacing the self-attention component of the standard Transformer with the ASSA mechanism described in Section 2.1. The query, key, and value matrices Q , K , and V from Equation (4) are directly substituted into Equations (A5) and (A6), and the attention output from Equation (A7) is fed into the remaining Transformer layers. In other words, the multi-head attention module in Figure 1 is replaced with ASSA, while the rest of the Transformer architecture remains unchanged. For simplicity, this architecture will be referred to as “ASSA” in the following text.

To explore the potential of ASSA–Transformer, while keeping other parts unchanged, we replace the position-wise FNN part (Equation (A4)) with a network that can capture and learn time-series features more deeply, which can be expressed as

$$\text{FNN}_{\text{Deep}}(x) = \text{LayerNorm}(x + W_3 \text{GELU}(W_2 \text{GELU}(W_1 x + b_1) + b_2)), \quad (5)$$

where b_1 and b_2 are bias terms, and GELU denotes the Gaussian Error Linear Unit activation function. In this design, W_1 first expands the dimensionality, W_2 subsequently reduces it, and W_3 projects the representation back to the original space, thereby forming a “bottle-neck” structure that expands and contracts the feature space. This architecture enhances the model’s representational capacity while maintaining the final output dimensionality consistent with the input for residual connections. The two successive GELU activations introduce nonlinearity, allowing the network to approximate highly complex functions. Finally, the incorporation of layer normalization ensures stable training and facilitates efficient information propagation through the residual pathway. This resulting architecture, i.e., the ASSA–Transformer equipped with FNN_{Deep} , will hereafter be referred to as “ASSA Deep” in subsequent discussions.

The final variant replaces the position-wise FNN component of the ASSA–Transformer with the FRFN described in Section 2.1. This architecture is hereafter referred to as the “ASSA FRFN” model. In our study, we employ the three architectures: “ASSA”, “ASSA Deep”, and “ASSA FRFN”, for PV forecasting and systematically compare their predictive accuracy against the baseline Transformer and other benchmark algorithms. Their structures and differences are shown in Figure 4, and Table 1 summarizes the structural and functional differences among these three architectures and the Transformer. From a computational complexity perspective, the four algorithms display a clear efficiency hierarchy. The baseline Transformer is dominated by the FFN layer, which accounts for about 73% of the total computation, while the ASSA–Transformer increases the complexity by only

3% through adaptive sparse attention. ASSA Deep enhances the feature extraction with a three-layer FFN but incurs a 76% increase in complexity, making it the most expensive variant. In contrast, ASSA FRFN introduces partial convolutions and gating, reducing the complexity by 18% compared with DeepFFN, with a convolutional cost only 1.2% of that of the standard FFN.

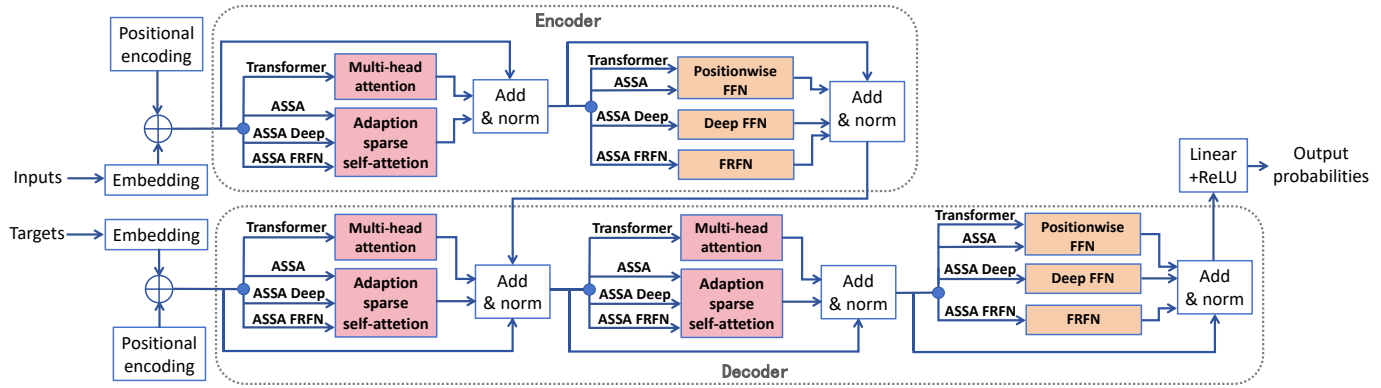


Figure 4. The architectures of the four algorithms, standard Transformer, ASSA, ASSA Deep, and ASSA FRFN, employ different attention mechanisms and feedforward neural network structures, as indicated by the arrows.

Table 1. Comparison of key characteristics among four algorithms. The dimension of each attention head is d_k . The hidden layer dimensions of the FFN $d_{ff} = 64$. The number of channels for partial convolution in the FRFN is $C_{\text{partial}} = 8$, with a kernel size of a one-dimensional partial CNN, $k = 3$.

Characteristic	Transformer	ASSA	ASSA Deep	ASSA FRFN
Attention Mechanism	Standard dot-product	ASSA	ASSA	ASSA
FFN Structure	FFN	FFN	Deep FFN	FRFN
Positional Encoding	None	Relative position bias	Relative position bias	Relative position bias
Feature Selection	None	Attention-level	Attention-level	Attention + Gating-level
Computational Complexity	$O(W^2 d_k)$	$O(W^2 d_k)$	$O(W d_{ff}^2)$	$O(W \cdot C_{\text{partial}} \cdot k)$

All algorithms ultimately share the same output layer, formulated as

$$\hat{Y} = \text{ReLU}\left(\text{Flatten}(H_{\text{dec}})W_f + b_f\right)W_y + b_y. \quad (6)$$

The output layer first takes the feature tensor $H_{\text{dec}} \in \mathbb{R}^{B \times W \times d_{\text{model}}}$ from the decoder and flattens it into a two-dimensional matrix $\in \mathbb{R}^{B \times (W \cdot d_{\text{model}})}$. Subsequently, two linear transformations are applied. The first employs weights $W_f \in \mathbb{R}^{(W \cdot d_{\text{model}}) \times d_{\text{ff}}}$ and bias $b_f \in \mathbb{R}^{d_{\text{ff}}}$ to project features into a latent space, followed by the ReLU activation function to introduce nonlinearity. The second transformation compresses features into the prediction space using $W_y \in \mathbb{R}^{d_{\text{ff}} \times T}$ and $b_y \in \mathbb{R}^T$, producing the final prediction $\hat{Y} \in \mathbb{R}^{B \times T}$.

This design preserves spatiotemporal feature information through flattening, enhances the representational power with nonlinear transformations, and provides a unified prediction interface for all algorithms. Since the input data is normalized prior to training, the prediction results must be denormalized to restore them to the original measurement scale. The final prediction $\hat{Y}$ can be denormalized as

$$\hat{Y}_{\text{inverse}} = \hat{Y} \cdot (X_{\text{max}} - X_{\text{min}}) + X_{\text{min}}. \quad (7)$$

Based on the theoretical framework of the ASSA-Transformer, the complete procedure employed in our study is summarized in Algorithm 1. We use Optuna to automatically optimize the hyperparameters of the standard Transformer, including the learning rate, dropout

rate, and embedding dimension, among others, and apply the resulting settings consistently across the Transformer-based ASSA, ASSA Deep, and ASSA FRFN architectures to ensure fair comparability. All models use the same input sequence length, prediction horizon, and standardized five-dimensional input features. The network architecture is uniformly configured with an embedding dimension of 32, a feedforward hidden dimension of 64, and an encoder–decoder structure composed of 4 attention heads and 3 Transformer layers. The training settings are also aligned, including a dropout ratio of 0.01, a learning rate of 0.001, a batch size of 32, and 50 training epochs, with input features normalized to the range $[0, 1]$. The loss function is defined as the Huber loss with $\delta = 0.5$, which behaves as the mean squared error (MSE) for deviations smaller than δ and approximates the mean absolute error (MAE) for larger deviations, thereby enhancing the robustness to outliers. This rigorous consistency guarantees that the observed performance differences arise solely from the core architectural variations, such as standard attention versus ASSA attention and standard FFN versus Deep FFN and FRFN, rather than from incidental hyperparameter choices. Consequently, the experimental setup provides a reliable foundation for evaluating the impact of architectural innovations.

Algorithm 1 Transformer with ASSA Variants for PV Time-Series Prediction.

```

1: Input: Time-series data  $X_{\text{input}} \in \mathbb{R}^{B \times W \times D}$ 
2: Parameters:
    • Embedding dim  $D_E = 32$ , attention dim  $d_k = 16$ 
    • FFN dim  $d_{\text{ff}} = 64$ ,  $C_{\text{partial}} = 8$ 
    • Weight matrices  $W_e, W^Q, W^K, W^V, W_f, W_y$ 
3: procedure EMBEDDING( $X_{\text{input}}$ )
4:    $E \leftarrow X_{\text{input}} W_e + b_e$ 
5:    $b_e(p, i) \leftarrow \begin{cases} \sin(p/10000^{2i/d_{\text{model}}}), & i \text{ even} \\ \cos(p/10000^{2i/d_{\text{model}}}), & i \text{ odd} \end{cases}$ 
6:   return  $E \in \mathbb{R}^{B \times W \times D_E}$ 
7: end procedure
8: procedure ATTENTIONPROJECTION( $E$ )
9:    $Q \leftarrow E W^Q, K \leftarrow E W^K, V \leftarrow E W^V$ 
10:  return  $(Q, K, V) \in \mathbb{R}^{B \times W \times d_k}$ 
11: end procedure
12: procedure ENCODERBLOCK( $E$ )
13:   $(Q, K, V) \leftarrow \text{AttentionProjection}(E)$ 
14:  for each layer  $l = 1$  to  $N$  do
15:     $\text{Attention} \leftarrow \begin{cases} \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V & \text{(Transformer)} \\ w_1 \text{Attention}_S + w_2 \text{Attention}_R & \text{(ASSA, ASSA Deep and FRFN)} \end{cases}$ 
16:     $E \leftarrow \text{LayerNorm}(E + \text{Dropout}(\text{Attention}))$ 
17:     $E \leftarrow \begin{cases} \text{ReLU}(E W_1) W_2 & \text{(Transformer and ASSA)} \\ \text{LayerNorm}(E + \text{DeepFFN}(E)) & \text{(ASSA Deep)} \\ \text{FRFN}(E) & \text{(ASSA FRFN)} \end{cases}$ 
18:  end for
19:  return encoded features
20: end procedure
21: procedure DECODERBLOCK( $E, H_{\text{enc}}$ )
22:  Similar to EncoderBlock with
23:    1. Masked attention
24:    2. Encoder–decoder attention using  $H_{\text{enc}}$ 
25:  return decoded features  $H_{\text{dec}} \in \mathbb{R}^{B \times W \times d_{\text{model}}}$ 
26: end procedure
27: procedure OUTPUTLAYER( $H_{\text{dec}}$ )
28:   $F \leftarrow \text{Flatten}(H_{\text{dec}}) \in \mathbb{R}^{B \times (W \cdot d_{\text{model}})}$ 
29:   $\hat{Y} \leftarrow \text{ReLU}(F W_f + b_f) W_y + b_y$ 
30:  return predictions  $\hat{Y} \in \mathbb{R}^{B \times T}$ 
31: end procedure
32:  $E \leftarrow \text{Embedding}(X_{\text{input}})$ 
33:  $H_{\text{enc}} \leftarrow \text{EncoderBlock}(E)$ 
34:  $H_{\text{dec}} \leftarrow \text{DecoderBlock}(E, H_{\text{enc}})$ 
35:  $\hat{Y} \leftarrow \text{OutputLayer}(H_{\text{dec}})$ 

```

To evaluate the performance of ASSA and its variants, we compare them with several baseline models, including Long Short-Term Memory (LSTM) [32], bidirectional LSTM (BiLSTM) [33], a Gated Recurrent Unit (GRU) [34], and a Temporal Convolutional Network (TCN) [35]. The LSTM model employs a two-layer architecture with 128 hidden units, trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The BiLSTM model uses the same configuration as LSTM, except that it has a two-layer bidirectional structure with 128 hidden units in each direction. The GRU model consists of a two-layer structure with 128 fully connected units, utilizes the ReLU activation function, applies a dropout rate of 0.2, and is trained with the Adam optimizer using a learning rate of 0.001 and a batch size of 32. The TCN model adopts a three-layer convolutional architecture with a kernel size of 3 and a dilation base of 2. Its fully connected layers are configured with 64, 128, and 72 units, and it also uses a dropout rate of 0.2, the Adam optimizer with a learning rate of 0.001, and a batch size of 32. All of the experiments are conducted on PyTorch 2.6.0 with Python 3.10.11, utilizing an NVIDIA GeForce RTX 4060 Ti (16 GB) GPU. The results of these algorithms were obtained in our previous study on long-term forecasting [36]. In this short-term forecasting study, the Transformer has been adapted to the ASSA–Transformer framework and optimized accordingly, so its results differ from the previous work.

3. Data Preprocessing

To ensure the reproducibility and rigor of our research, this study utilizes a publicly available dataset from the 5.4 kW amorphous silicon photovoltaic (PV) system at the Calyxo power plant, located within the Desert Knowledge Australia Solar Centre (DKASC) in Alice Springs, Australia. The DL models will be trained to predict PV power generation using four key environmental variables: ambient temperature ($^{\circ}\text{C}$), relative humidity (%), global horizontal radiation (W/m^2), and diffuse horizontal radiation (W/m^2). These variables are selected based on their statistically significant correlations with PV power output, as evidenced by Pearson's correlation coefficients of 0.28, -0.38 , 0.97, and 0.25, respectively. The validity of these input variables has been further supported by prior research. The selection of highly correlated features enhances the model's ability to learn underlying physical relationships while optimizing computational efficiency. The dataset comprises recordings at 5 min intervals. To account for negligible nocturnal PV generation, the analysis is restricted to daylight hours (07:00–19:00). Data preprocessing includes rigorous quality control measures: days with >10 missing values are excluded, while minor gaps (<10 missing values) are addressed via linear interpolation. After post-processing, the total missing data accounts for merely 2.21% of the dataset (below the 3% threshold), thus preserving data integrity for robust deep learning applications.

For temporal validation, the model is trained on data from 2020 to 2022 and evaluated on the 2023 test set. This chronological partitioning is designed to simulate a realistic deployment scenario, training the model on historical data to predict the future performance and thereby providing a pragmatic assessment of its generalizability. The three-year training window from 2020 to 2022 is selected to capture a sufficiently diverse range of meteorological conditions, including seasonal patterns as well as anomalous events such as dust storms and heat waves. This design reduces the likelihood of overfitting to the unique climate characteristics of a single year while also offering a rigorous and practical out-of-sample evaluation. The data of 2023 is reserved as a holdout set representing the unknown future, effectively testing the model's ability to extrapolate from past patterns without assuming stationarity in the weather time series. Moreover, because our PV forecasts focus on short-term horizons (5 min to 1 h ahead), which are more strongly influenced by

short-term anomalies than by long-term annual variability, the chosen training and test partitioning is both reasonable and appropriate.

The characteristics of the DKASC dataset guide the design of our modeling architecture. Its high temporal resolution enables the capture of rapid fluctuations in desert climates, while the prevalence of high-frequency noise necessitates robust feature selection and filtering, which we address with the ASSA mechanism. To better approximate the nonlinear relationships between environmental variables and power output, we extend the standard FNN into a deeper bottleneck structure (Deep FNN). In addition, we explore the FRFN with gating and partial convolution to efficiently emphasize informative temporal features. Overall, these architectural choices directly respond to the spatiotemporal complexity and noisy nature of the dataset.

While the DKASC dataset provides high-quality, high-resolution data well suited to method development and benchmarking, an important limitation of this study must be acknowledged. The model is trained and evaluated using data from a single geographic location in Alice Springs, Australia, which is characterized by a desert climate. This focus inherently constrains the generalizability of the findings to regions with substantially different climate regimes, such as tropical zones with high humidity and frequent convective cloud formation or temperate maritime climates dominated by diffuse solar radiation. Consequently, the demonstrated advantages of the ASSA mechanism are validated primarily within the context of desert-specific conditions, including high irradiance, rapid fluctuations from clear to cloud-covered skies, and low aerosol content. Nonetheless, the core strengths of ASSA, namely adaptive noise suppression and salient feature selection, constitute fundamental architectural improvements that hold promise for time-series forecasting across diverse domains. However, publicly available PV datasets with high-frequency measurements and detailed environmental attributes, such as DKASC, remain scarce. Future work will focus on multi-site validation across a broader range of climatic and geographic settings to assess the general applicability and robustness of the proposed framework.

To assess the predictive performance of different models, we employ three widely used metrics: the coefficient of determination R^2 , the mean absolute error (MAE), and the root mean square error (RMSE), defined as follows:

$$\begin{aligned} R^2 &:= 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y} - y_i)^2}, \\ \text{MAE} &:= \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|, \\ \text{RMSE} &:= \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \end{aligned} \quad (8)$$

where $\hat{y}_i$ denotes the predicted PV power, and y_i represents the corresponding observed value. Higher R^2 values indicate predictions that closely match the actual data, whereas lower MAE and RMSE values correspond to better predictive accuracy.

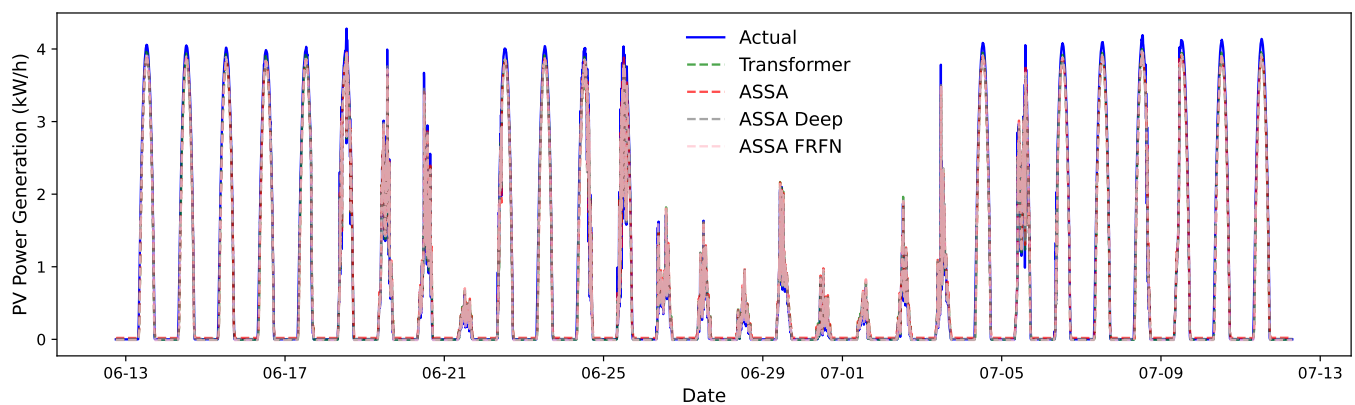
4. Results

4.1. Five-Minute-Ahead Forecasting

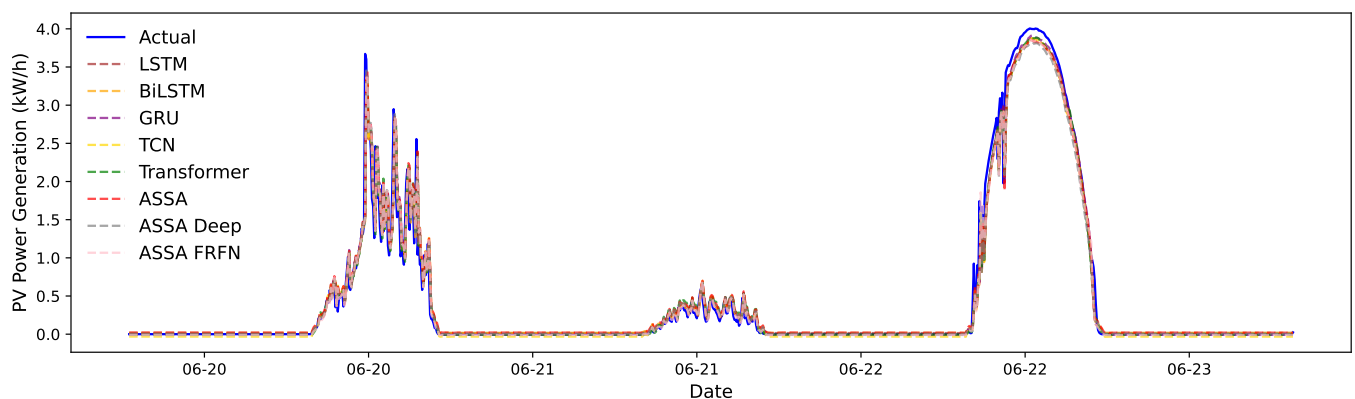
Five-minute-ahead PV power generation forecasting is crucial for improving power system stability and facilitating the efficient integration of renewable energy. PV generation exhibits intrinsic variability due to meteorological factors, particularly fluctuations in irradiance, which cause rapid short-term changes in power output. High-resolution forecasts at 5 min intervals enable real-time monitoring of these dynamics, allowing grid operators to implement timely dispatch adjustments and reduce instability arising from

sudden power imbalances. Accurate short-term forecasts enhance operational efficiency by optimizing the use of flexible resources, such as energy storage systems, and reducing dependence on the reserve capacity. This precision minimizes the risks of frequency deviations and involuntary load shedding caused by forecast errors. Furthermore, an improved forecast accuracy assists PV plant operators in lowering deviation penalties in electricity markets and formulating economically optimal bidding strategies, thereby maximizing operational profitability.

Figure 5a presents a comparative time-series analysis of 5 min ahead PV power generation forecasts produced by the four models described in Section 2.2 (Transformer, ASSA, ASSA Deep, and ASSA FRFN) against actual generation data from 13 Juneto 13 July 2023. The selected evaluation period includes diverse meteorological conditions, such as clear-sky, precipitation, and overcast days, as well as transient weather fluctuations. The visualization demonstrates the predictive performance of each model across varying weather patterns. All models closely align with the measured power output, successfully capturing the rapid short-term variability in PV generation caused by environmental factors. The predicted trajectories closely follow the actual power curves, particularly during periods of abrupt irradiance changes.



(a)



(b)

Figure 5. (a) Five-minute-ahead PV power generation forecasting of different algorithms from 13 June to 13 July 2023; (b) a comparison of five-minute-ahead PV power generation forecasting for different algorithms from 20 to 22 June 2023.

To provide a comprehensive assessment of prediction accuracy and facilitate comparison with other deep learning algorithms, Figure 5b shows the 5 min PV power generation forecasts from 20 to 23 June 2023 generated using the LSTM, BiLSTM, GRU, and TCN

models, alongside the actual generation data. This evaluation period includes diverse weather conditions, encompassing pre-rain cloudy days, rainy days, and post-rain sunny days. The results indicate systematic prediction biases, including both underestimation and overestimation, even when using advanced architectures such as Transformer, ASSA, ASSA Deep, and ASSA FRFN.

For a more rigorous comparison, Table 2 presents the performance evaluation metrics R^2 , MAE, and RMSE of each model for 5 min ahead PV power generation forecasting. The Transformer demonstrates a superior performance in the error metrics, achieving the lowest MAE (0.1454) and RMSE (0.3264) among all compared algorithms. This superior accuracy in ultra-short-term forecasting is critical for real-time dispatch and Automatic Generation Control (AGC), as it enables grid operators to determine more precise power set-points for resource balancing, thereby reducing costly frequency regulation requirements and minimizing the risk of deploying excessive spinning reserves. While its $R^2 = 0.9582$ approaches the maximum value of 0.9583 attained by ASSA, the latter shows only marginal improvement in this metric. Notably, ASSA exhibits slightly higher MAE (0.1475) and RMSE (0.3265) values compared to the Transformer. The ASSA FRFN variant achieves a comparable performance to the Transformer in terms of R^2 (0.9582) but shows inferior results for both the MAE (0.1483) and RMSE (0.3266). In contrast, ASSA Deep exhibits the weakest performance among the ASSA variants, with the lowest R^2 (0.9568) and highest RMSE (0.3319), suggesting that increased model depth may adversely affect prediction accuracy. These results indicate that while the ASSA variants remain competitive, the Transformer consistently demonstrates the most balanced and robust forecasting performance.

Table 2. The comparison of 5 min ahead PV power generation forecasting for different algorithms. The optimal value in each column is highlighted in bold.

Model	R^2	MAE	RMSE
LSTM	0.9572	0.1561	0.3362
BiLSTM	0.9581	0.1514	0.3301
GRU	0.9573	0.1562	0.3308
TCN	0.9578	0.1483	0.3294
Transformer	0.9582	0.1454	0.3264
ASSA	0.9583	0.1475	0.3265
ASSA Deep	0.9568	0.1513	0.3319
ASSA FRFN	0.9582	0.1483	0.3266

While the standard Transformer achieves the best MAE and RMSE for the 5 min ahead forecasting, it is important to note that the ultra-short-term prediction is predominantly influenced by the most recent temporal patterns and exhibits minimal noise accumulation. In such scenarios, the dense attention mechanism of the standard Transformer effectively captures immediate dependencies without being hindered by redundancy. However, as shown in subsequent sections, as the forecasting horizon extends, the ability of the ASSA mechanism to filter out irrelevant interactions and suppress noise becomes increasingly critical, leading to its superior performance in medium-term forecasts, e.g., 1 h ahead.

4.2. One-Hour-Ahead Forecasting

Compared to 5 min ahead forecasts, 1 h ahead PV power generation forecasts demonstrate greater adaptability to power system scheduling and market transactions. Their primary advantage stems from their alignment with the critical timescales of power system operations. Specifically, 1 h ahead forecasts can effectively integrate the correction cycles of numerical weather prediction (NWP) models, which typically operate at an approximate

resolution of 1 h. This capability provides essential decision-making support for economic dispatch and electricity market transactions.

Figure 6 presents the 1 h ahead PV power generation forecasts produced by the deep learning models (LSTM, BiLSTM, GRU, TCN, Transformer, and the ASSA variants) in comparison with the actual generation data from 20–23 June 2023. Relative to the 5 min ahead forecasts shown in Figure 5, the 1 h forecasts exhibit a marked decline in accuracy, particularly in capturing rapid fluctuations during rainy conditions (e.g., the rainfall event on 21 June). Nevertheless, the ASSA-based models, especially ASSA and ASSA Deep, remain more closely aligned with the observed power curve, demonstrating stronger robustness in tracking generation trends across weather transitions. By contrast, conventional models such as LSTM and the TCN display pronounced lag or overshooting under variable weather, underscoring the limitations of recurrent and convolutional structures in modeling 1 h dependencies under complex meteorological conditions. These findings confirm that the ASSA mechanism enhances temporal feature selection and noise suppression, thereby improving forecasting stability in challenging weather scenarios.

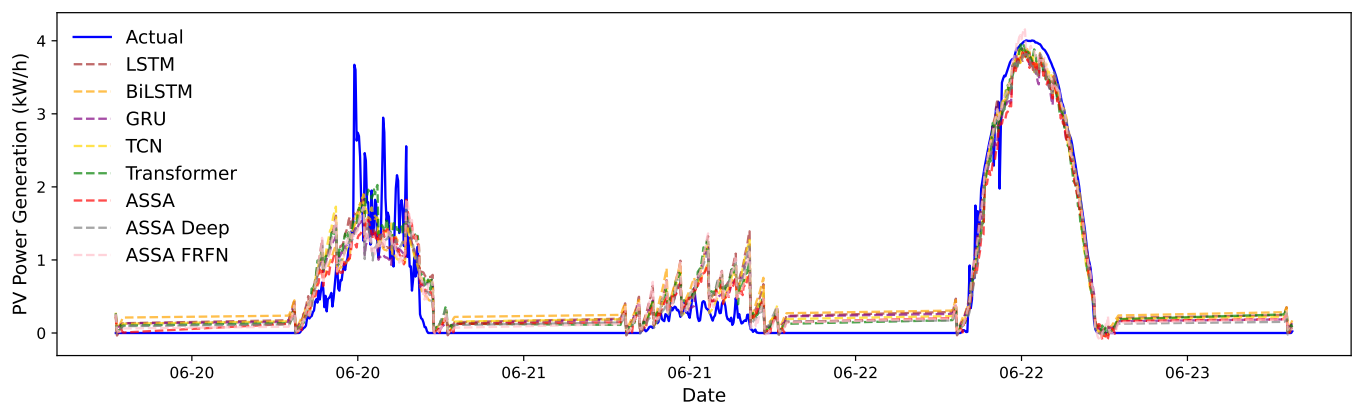


Figure 6. The comparison of 1 h ahead PV power generation forecasting for different algorithms from 20 to 22 June 2023.

Table 3 presents a comprehensive comparison of forecasting algorithms, where the ASSA model demonstrates a superior performance across all evaluation metrics. With an R^2 of 0.9115 (the highest among all models), an MAE of 0.2659 (4.4% lower than that of the Transformer), and an RMSE of 0.4752 (a 0.3% improvement over the Transformer), ASSA establishes itself as the most accurate predictor. This marginal improvement at the 1 h horizon has practical implications for electricity market operations and unit commitment, as it enables PV plant operators to make more accurate bids in day-ahead and intraday markets, thereby reducing financial penalties for generation deviations, while allowing system operators to schedule conventional generators with greater confidence, resulting in more economical and secure grid operations. While the Transformer baseline shows strong results ($R^2 = 0.9110$), the ASSA–Transformer variants collectively outperform traditional architectures, with ASSA Deep maintaining competitive accuracy (MAE = 0.2662). Notably, ASSA achieves the best balance between precision (lowest MAE/RMSE) and explanatory power (highest R^2), suggesting its enhanced capability in capturing PV power generation patterns. These results highlight the effectiveness of ASSA for 1 h ahead forecasting tasks.

Statistical tests show highly significant performance differences (p -values < 0.001) between ASSA and several mainstream models, including LSTM, GRU, TCN, and the baseline Transformer. This demonstrates the reproducibility and statistical robustness of ASSA’s predictions, demonstrating that its performance is not accidental. Furthermore, when comparing ASSA to the current state-of-the-art Transformer, the R^2 value is slightly

higher, indicating a “small” effect size. Combined with the extremely high statistical significance, this demonstrates that ASSA has the potential to compete with leading models, with a performance gap that is both statistically and practically significant.

Table 3. The comparison of 1 h ahead PV power generation forecasting for different algorithms. The optimal value in each column is highlighted in bold.

Model	R ²	MAE	RMSE
LSTM	0.9067	0.2982	0.4961
BiLSTM	0.9078	0.2852	0.4890
GRU	0.9060	0.2984	0.4975
TCN	0.9047	0.3101	0.5071
Transformer	0.9110	0.2719	0.4764
ASSA	0.9115	0.2659	0.4752
ASSA Deep	0.9091	0.2662	0.4814
ASSA FRFN	0.9055	0.2708	0.4909

4.3. Analysis of Short-Term Forecast Accuracy Trends

To comprehensively evaluate and compare the short-term forecasting performance of different models and to assess the effectiveness of the ASSA mechanism, we analyze the R², MAE, and RMSE of the Transformer, ASSA, ASSA FRFN, and ASSA Deep across forecast horizons ranging from 5 min ahead to 1 h ahead, using 5 min intervals. Table 4 presents a detailed frequency-dependent comparison of the R² scores among ASSA variants and Transformer models. The ASSA framework demonstrates consistent dominance, achieving the highest R² values in 7 out of 12 frequency intervals (5 min, 10 min, 20 min, 30 min, 35 min, 45 min, and 60 min), which establishes its strongest advantage at the 60 min forecasting horizon. The Transformer model leads only at 25 min (0.9352) and 40 min (0.9205), suggesting its advantage diminishes at higher forecasting frequencies. The data reveals an expected degradation in R² values with increasing time horizons, from 0.9583 (5 min) to 0.9115 (60 min) for ASSA, with all models maintaining an R² > 0.90 throughout, indicating a robust predictive capability across all tested frequencies.

Table 4. The trends in short-term forecasting accuracy measured by R² for the Transformer, ASSA, ASSA Deep, and ASSA FRFN. The optimal value in each step (row) is highlighted in bold.

Frequencies	R ²			
	ASSA	ASSA FRFN	ASSA Deep	Transformer
5 min	0.9583	0.9578	0.9568	0.9582
10 min	0.9481	0.9474	0.9475	0.9477
15 min	0.9436	0.9438	0.9430	0.9436
20 min	0.9373	0.9367	0.9358	0.9359
25 min	0.9321	0.9334	0.9335	0.9352
30 min	0.9282	0.9260	0.9282	0.9280
35 min	0.9249	0.9217	0.9247	0.9229
40 min	0.9204	0.9204	0.9182	0.9205
45 min	0.9184	0.9172	0.9161	0.9172
50 min	0.9146	0.9160	0.9149	0.9169
55 min	0.9103	0.9110	0.9122	0.9118
60 min	0.9115	0.9055	0.9091	0.9110

A systematic evaluation of the MAE is given in Table 5, across different forecasting horizons for ASSA variants and Transformer models. The ASSA architecture demonstrates a superior performance, achieving the lowest MAE values in 9 out of 12 frequency intervals. The standard Transformer model excels at the shortest 5 min interval (0.1454,

best among all), where recent temporal correlations are strong and noise is less impactful. However, ASSA quickly establishes dominance from 10 min onward as the need for robust feature selection and noise resilience grows. The ASSA FRFN variant shows a competitive performance at 20 min (0.1928, best result), while ASSA Deep achieves the optimal results at 30 min (0.2160). The data in Table 5 reveals a clear trend of increasing MAE values with longer forecasting horizons across all models, with ASSA consistently maintaining the lowest error growth rate.

Table 5. The trends in short-term forecasting accuracy measured using the MAE for the Transformer, ASSA, ASSA Deep, and ASSA FRFN. The optimal value in each step (row) is highlighted in bold.

Frequencies	MAE			
	ASSA	ASSA FRFN	ASSA Deep	Transformer
5 min	0.1475	0.1515	0.1513	0.1454
10 min	0.1557	0.1676	0.1721	0.1596
15 min	0.1734	0.1737	0.1771	0.1755
20 min	0.1929	0.1928	0.1977	0.1982
25 min	0.2017	0.2089	0.2042	0.2046
30 min	0.2194	0.2223	0.2160	0.2236
35 min	0.2171	0.2396	0.2260	0.2224
40 min	0.2336	0.2419	0.2442	0.2451
45 min	0.2472	0.2478	0.2536	0.2501
50 min	0.2518	0.2526	0.2543	0.2573
55 min	0.2602	0.2655	0.2565	0.2660
60 min	0.2659	0.2708	0.2662	0.2719

Table 6 shows a comprehensive evaluation of the RMSE performance across various forecasting horizons. Similar to previous observations, the ASSA model exhibits superior predictive accuracy, achieving the lowest RMSE values in eight out of twelve forecast horizons and demonstrating the greatest advantage at the 60 min horizon with an RMSE of 0.4752, surpassing both the Transformer baseline, which has an RMSE of 0.4764, and the other ASSA variants. While the Transformer model shows a competitive performance at shorter intervals (5 min: 0.3264, best result; 15 min: 0.3792), ASSA consistently delivers better results for short-to-medium-term predictions (20 to 60 min). The ASSA Deep variant demonstrates particular strength at 30 min (0.4279, best among all) and 55 min (0.4733), while the ASSA FRFN shows a competitive performance at 50 min (0.4628). The data reveals a clear progression of increasing RMSE values with longer forecasting horizons across all models, with ASSA maintaining the most stable error growth from 0.3265 (5 min) to 0.4752 (60 min).

A comprehensive evaluation using R^2 , MAE, and RMSE metrics demonstrates that the ASSA framework exhibits significant advantages in PV power forecasting. Regarding prediction accuracy (R^2), ASSA achieves a superior performance at 7 out of 12 time frequencies. In terms of the MAE, ASSA maintains the lowest error values at 9 of 12 intervals. The RMSE results further confirm ASSA's stability, with the framework attaining minimal error values at 8 of 12 frequencies. Notably, while the Transformer model delivers the optimal performance for ultra-short-term (5 min) forecasts, ASSA establishes its dominance beginning with medium-term (10 min) predictions. Furthermore, ASSA shows the most gradual error growth rate as the forecast duration increases, with the MAE and RMSE increasing by only 80.2% and 45.5% respectively—the lowest growth rates among all evaluated models. The ASSA variants exhibit specialized strengths: ASSA Deep demonstrates an exceptional performance for 30 min forecasts (the lowest RMSE = 0.4279), suggesting that its deeper FFN is particularly effective at capturing the medium-term temporal patterns that dominate

this forecasting horizon, while ASSA FRFN shows particular competitiveness in specific frequency bands (e.g., 50 min forecasts). All models maintain a robust predictive capability ($R^2 > 0.90$), verifying the reliability of DL algorithms for PV forecasting. Importantly, the ASSA–Transformer framework delivers the best overall performance by effectively balancing short-term accuracy with long-term stability. The consistent accuracy of ASSA across multiple time horizons, from 5 to 60 min, provides a unified and reliable data stream for various grid control layers. This simplifies the coordination between short-term balancing actions and medium-term scheduling, ensuring that decisions made for AGC are consistent with the economic dispatch plan, thereby enhancing overall operational efficiency.

Table 6. The trends in short-term forecasting accuracy measured using the RMSE for the Transformer, ASSA, ASSA Deep, and ASSA FRFN. The optimal value in each step (row) is highlighted in bold.

Frequencies	RMSE			
	ASSA	ASSA FRFN	ASSA Deep	Transformer
5 min	0.3265	0.3281	0.3319	0.3264
10 min	0.3639	0.3662	0.3658	0.3653
15 min	0.3794	0.3788	0.3812	0.3792
20 min	0.3999	0.4017	0.4047	0.4042
25 min	0.4106	0.4122	0.4118	0.4162
30 min	0.4283	0.4343	0.4279	0.4285
35 min	0.4376	0.4468	0.4383	0.4434
40 min	0.4505	0.4505	0.4567	0.4505
45 min	0.4562	0.4595	0.4575	0.4596
50 min	0.4666	0.4628	0.4659	0.4604
55 min	0.4784	0.4766	0.4733	0.4744
60 min	0.4752	0.4909	0.4814	0.4764

4.4. Forecasting in Different Weather Conditions

Weather dependency is a critical factor in PV power generation forecasting. In this study, we systematically evaluate the forecasting performance of the DL models under three representative weather conditions: clear, rainy, and cloudy. As summarized in the Table 7, the comparison of R^2 , MAE, and RMSE metrics for the models within a 1 h ahead forecasting horizon highlights the superior adaptability of the models across the weather conditions. Here, we focus on the 1 h ahead prediction because it aligns more closely with the dispatch cycle of the provincial power grid’s AGC system, thereby providing direct support for economic dispatch and unit commitment decisions. Compared with 5 min ahead forecasts, which are prone to transient cloud-induced fluctuations, 1 h ahead forecasting more effectively captures the evolving dynamics of weather systems, mitigating the cumulative errors inherent in shorter intervals while achieving an optimal balance between computational resource requirements and forecasting accuracy.

For sunny day forecasts, the BiLSTM model achieved the best overall performance, with the highest R^2 (0.9420) and the lowest RMSE (0.3748), highlighting its strong capability in time-series modeling under stable irradiance conditions. The ASSA FRFN obtained the lowest MAE (0.2159), indicating its effectiveness in accurately capturing minor power fluctuations on clear days. Notably, the ASSA-FRFN variant achieved the lowest MAE (0.2159), indicating a superior performance in capturing small power fluctuations under sunny conditions. This improvement can be attributed to its gated feature refinement mechanism, which effectively processes the smooth, high-irradiance sequences characteristic of sunny days by enhancing relevant features while suppressing residual noise. All models achieved an R^2 exceeding 0.93, confirming the relative ease of forecasting in high-irradiance envi-

ronments. Furthermore, the BiLSTM model outperformed the next-best performer, ASSA Deep ($R^2 = 0.9408$), by approximately 0.3%.

Table 7. Model performance comparison of 1 h ahead PV prediction across weather conditions: sunny, rainy, and cloudy. The optimal value of each metric (column) is highlighted in bold.

Model	Sunny			Rainy			Cloudy		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
LSTM	0.9389	0.2337	0.3847	0.7258	0.4658	0.7036	0.7335	0.4844	0.7396
BiLSTM	0.9420	0.2194	0.3748	0.7324	0.4549	0.6951	0.7344	0.4749	0.7384
GRU	0.9378	0.2381	0.3881	0.7367	0.4464	0.6894	0.7355	0.4721	0.7368
TCN	0.9340	0.2503	0.4000	0.7170	0.4590	0.7148	0.7332	0.4796	0.7401
Transformer	0.9401	0.2394	0.3809	0.7370	0.4729	0.6890	0.7434	0.4871	0.7258
ASSA	0.9392	0.2376	0.3838	0.7463	0.4416	0.6767	0.7489	0.4702	0.7181
ASSA Deep	0.9408	0.2234	0.3788	0.7354	0.4574	0.6912	0.7486	0.4650	0.7184
ASSA FRFN	0.9405	0.2159	0.3798	0.7259	0.4586	0.7034	0.7313	0.4720	0.7427

The ASSA model demonstrated overall superiority in rainy scenarios, achieving the best R^2 (0.7463), MAE (0.4416), and RMSE (0.6767), representing a 2.05% improvement in R^2 compared with the traditional LSTM. By leveraging its adaptive feature selection mechanism, ASSA effectively mitigated high-frequency noise induced by rainfall. Among the remaining models, the GRU achieved the second-best performance in the MAE (0.4464) and RMSE (0.6894); however, ASSA still yielded notable error reductions of 1.1% (MAE) and 1.9% (RMSE).

Under cloudy conditions, the ASSA series also exhibits a dominant performance: ASSA achieves the best R^2 (0.7489) and RMSE (0.7181), while ASSA Deep obtains the lowest MAE (0.4650). This divergence highlights the ASSA architecture's unique ability to handle rapidly varying cloud cover. Specifically, the basic ASSA excelled at capturing overall trends (with R^2 improved by 1.54% compared with that of the Transformer), whereas ASSA Deep is more effective at modeling instantaneous fluctuations (with the MAE reduced by 1.1% compared with that of basic ASSA). It is noteworthy that forecasting under cloudy conditions remains substantially more challenging than under clear skies: the best R^2 (0.7489) is 19.31% lower than the best clear-sky result, underscoring the critical role of weather complexity in determining forecasting accuracy.

The comparison results across various weather scenarios demonstrate that ASSA and its derivatives provide significant advantages in PV power forecasting. Under clear-sky conditions, all models achieve a high predictive accuracy, with the ASSA algorithms slightly outperforming LSTM. In rainy scenarios, the ASSA model effectively suppresses high-frequency noise induced by rainfall through its adaptive feature selection mechanism, yielding substantial improvements in the R^2 , MAE, and RMSE compared with those for traditional LSTM and GRU models. Under cloudy conditions, the ASSA family of algorithms maintains its superiority: the basic ASSA is more effective at capturing overall trends, while ASSA Deep excels at modeling transient fluctuations. This complementarity highlights the unique capability of the ASSA architecture to handle rapidly changing cloud cover, demonstrating superior adaptability and robustness under complex meteorological conditions. By achieving a balanced representation between long-term trend characterization and short-term fluctuation modeling while maintaining a high forecasting accuracy, ASSA and its variants show strong potential for practical applications in power grid scheduling and economic operations. Their robustness under volatile and rainy weather conditions is particularly vital for Grid Reliability and Contingency Planning, as accurate predictions during these periods enable system operators to secure additional reserve capacity in

advance, thereby reducing the risk of voltage instability and potential overloads caused by sudden large-scale reductions in PV generation.

5. Discussion

5.1. Stability and Efficiency of the Algorithms

To ensure the comparability of the algorithms, the number of training epochs is fixed at 50. Figure 7 presents the one-hour-ahead loss curves of all algorithmic architectures employed in the previous section. As shown, the loss values of all models converge by the 50th epoch, indicating stable training and the effectiveness of the algorithms.

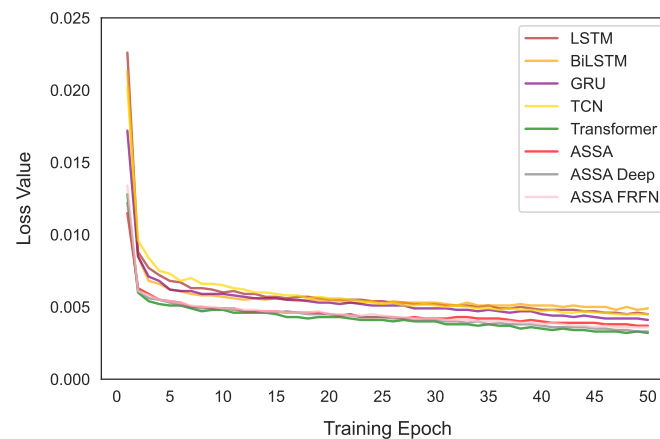


Figure 7. Loss values of all algorithms.

Table 8 summarizes the computational resource requirements of different deep learning models for 1 h ahead PV power forecasting. The results indicate a clear trade-off between model complexity and forecast accuracy, which is critical for practical deployment. Traditional recurrent architectures such as LSTM, BiLSTM, and the GRU require relatively short training times (130–188 s) but deliver a notably lower accuracy under variable weather conditions compared with attention-based models. The Transformer, despite a longer training time of 327 s, achieves substantial improvements in accuracy, particularly for the ultra-short-term horizon, highlighting the effectiveness of self-attention in capturing complex spatiotemporal dependencies.

Table 8. Computational resource requirements for 1 h ahead PV forecasting with different algorithms.

Model	Training Time	Peak Memory Usage	Average CPU Usage
LSTM	130 s	1031.44 MB	16.20%
BiLSTM	146 s	1036.15 MB	21.45%
GRU	188 s	1031.21 MB	19.92%
TCN	228 s	1107.80 MB	17.96%
Transformer	327 s	1036.17 MB	21.26%
ASSA	331 s	1291.10 MB	22.30%
ASSA Deep	512 s	1280.35 MB	22.34%
ASSA FRFN	408 s	1311.58 MB	28.99%

The ASSA model trains in 331 s, nearly matching the efficiency of the Transformer, yet consistently surpasses it in medium-term forecasting. This modest increase in training time is accompanied by a slightly higher memory usage (1291.10 MB vs. 1036.17 MB) and CPU utilization (22.30% vs. 21.26%), demonstrating that the adaptive sparse self-attention mechanism improves robustness and generalization at a minimal computational cost, even under challenging meteorological conditions such as rain and cloud cover.

In contrast, the more complex variants, ASSA Deep and the ASSA FRFN, are substantially more resource-intensive. ASSA Deep requires 512 s of training (56.6% longer than the Transformer), while the ASSA FRFN requires 408 s (24.8% longer). Both models consume more memory, with the ASSA FRFN reaching 1311.58 MB, and show higher CPU utilization, particularly the ASSA FRFN at 28.99%. However, their accuracy gains are limited, indicating diminishing returns with increasing architectural complexity.

From a deployment perspective, the standard Transformer and ASSA models achieve a favorable balance between accuracy and computational cost, making them more suitable for real-world applications where efficiency and scalability are critical. In contrast, the higher resource demands of ASSA Deep and the ASSA FRFN may restrict their use in hardware-constrained environments, edge computing scenarios, or contexts requiring frequent retraining. Overall, while ASSA enhances the performance with minimal overhead, further research should focus on improving the efficiency of more complex variants, for example, through model compression, quantization, or dynamic computation, to ensure practical applicability in operational settings.

Next, we evaluate the statistical confidence of our results. To ensure robustness, all reported performance metrics, including R^2 , MAE, and RMSE, represent the mean values computed over 10 independent experimental trials. Given that the Transformer-based ASSA variants investigated in this study exhibit similar architectures and confidence levels, we focus our analysis on the ASSA algorithm as a representative case. Table 9 summarizes the aggregated performance metrics across all experimental repetitions, demonstrating both the high predictive accuracy and stability of the proposed model. The standard deviation of R^2 is ± 0.0015 , indicating an excellent model fit with minimal variation. The MAE and RMSE are 0.2659 ± 0.0069 and 0.4825 ± 0.0045 , respectively, reflecting stable error control and low dispersion, with coefficients of variation below 2.6%. Moreover, the 95% confidence intervals for all metrics are narrow, and the corresponding means are located near the centers of the intervals, suggesting a symmetrical and unbiased distribution of results. Overall, the small standard deviations and concentrated confidence intervals demonstrate that the ASSA algorithm achieves a consistent performance across repeated experiments, thereby confirming its statistical robustness and reliability.

Table 9. A statistical summary of the ASSA performance metrics across 10 experimental runs for 1 h ahead PV forecasting.

Metrics	Mean	Standard Deviation	Confidence Interval (95%)
R^2	0.9115	0.0015	[0.9104, 0.9126]
MAE	0.2659	0.0069	[0.2613, 0.2705]
RMSE	0.4825	0.0045	[0.4722, 0.4882]

5.2. Impact of Prediction Error on PV Grid Operation

Our experimental results on PV power generation forecasting demonstrate that the enhanced accuracy of the proposed ASSA–Transformer model provides substantial practical value for grid operators. Reducing forecast errors directly improves operational efficiency and lowers costs. In power systems with a growing share of renewable energy, accurate forecasts are essential for stabilizing grid operations and optimizing dispatch decisions.

For ultra-short-term forecasts (e.g., 5 min ahead), the standard Transformer achieves the lowest error among all compared models, reducing the MAE to 0.1454 and the RMSE to 0.3264. This improvement allows grid dispatch systems to better capture instantaneous fluctuations in PV output, thereby enabling more flexible adjustments of reserve capacity and rapid-response resources. A lower forecast error also reduces the risk of frequency deviations and imbalances. In regions with high renewable penetration, such refined

forecasts help decrease the reliance on traditional fast-ramping generators and enhance the economic efficiency of system operations.

At the one-hour horizon, the ASSA model exhibits even greater robustness, achieving a MAE of 0.2659 (a 2.2% improvement over the standard Transformer) and an RMSE of 0.4752, outperforming all baseline models. Forecasts at this timescale are closely tied to market transactions and day-ahead scheduling. Improved accuracy enables operators to submit more precise market bids, reduce penalties from forecast bias, and enhance the reliability of PV output assessments, thereby supporting more efficient unit commitment and economic dispatch.

The advantages of the ASSA model are particularly evident in rainy conditions, where its R^2 reaches 0.7463 of a 2.05% improvement over LSTM and its MAE decreases to 0.4416 of a 5.2% improvement. This robustness under adverse weather provides reliable data support for grid dispatch, effectively mitigating the voltage instability and equipment overload caused by sudden fluctuations in power output.

From the perspective of computational efficiency, the ASSA model requires only 331 s to train, which is comparable to the standard Transformer (327 s), with a peak memory usage of 1291.10 MB. This balance of accuracy and efficiency makes the model suitable for real-world deployment, offering grid operators a practical and effective decision-support tool.

From a long-term perspective, an enhanced forecasting capability will facilitate large-scale grid integration and market-oriented operation of PV power. By reducing errors, the ASSA–Transformer not only advances the technical accuracy of PV forecasting but also delivers operational benefits, including improved dispatch, cost reduction, and risk mitigation. It thus provides an essential tool for developing a resilient and intelligent new power system.

6. Conclusions

This study introduces Transformer variants based on the adaptive sparse self-attention (ASSA) mechanism to systematically evaluate the impact of architectural enhancements on short-term PV power forecasting. These variants adopt a dual-branch attention structure that combines sparse and dense paths with adaptive weighting, thereby suppressing noise while preserving spatiotemporal dependencies. Additional components such as a deep feedforward network and a feature refinement feedforward network (FRFN) further strengthen feature extraction and refinement.

The ASSA-based framework outperforms conventional models including LSTM, the GRU, the TCN, and the standard Transformer, particularly in 1 h ahead forecasts and under challenging conditions such as overcast or rainy weather. While the baseline Transformer remains competitive in ultra-short-term forecasting (e.g., 5 min ahead), ASSA demonstrates superior generalization and stability as the horizon lengthens, underscoring its practical value for grid dispatch and power market operations. Moreover, compared to the standard Transformer, the ASSA mechanism achieves adaptive sparsity and noise suppression without incurring a significant computational overhead, thereby ensuring scalability and suitability for real-world deployment.

Despite these advances, several limitations remain. The models were trained and evaluated on a single dataset from a desert climate region (Alice Springs, Australia), which may constrain their applicability to other climates. The forecast horizon was restricted to one hour ahead, and the performance was not validated across multiple PV plants or spatial settings. Future work should therefore focus on multi-site and cross-climate validation, integration with numerical weather prediction to extend horizons, and probabilistic forecasting to capture uncertainty. Applying these models to large-scale PV and virtual

power plants, together with transfer learning for rapid adaptation to new sites, represents a promising direction for further research.

Author Contributions: Conceptualization: X.Z. and F.L.; methodology: X.Z., F.L., and Y.W.; software: F.L.; validation: F.L. and Y.W.; formal analysis: F.L. and Y.W.; writing—original draft preparation: F.L. and M.L.; writing—review and editing: X.Z., F.L., M.L., and Y.W.; supervision: F.L.; project administration: F.L.; funding acquisition: X.Z., F.L., M.L., and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Yunnan Fundamental Research Projects (Grant No. 202401AU070125), the Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities' Association (Grant No. 202301BA070001-114), Yunnan Provincial Department of Education Science Research Fund Project (Grant No. 2025J0942), 2025 Self-funded Science and Technology Projects of Chuxiong Prefecture (Grant No. cxzc2025008), Chuxiong Normal University Doctoral Research Initiation Fund Project (Grant No. BSQD2407) and Dongying Science Development Fund (Grant No. DJB2023015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting this study's findings is available from the DKA Solar Centre and can be accessed online at <https://dkasolarcentre.com.au/download?location=alice-springs> (accessed on 11 January 2025).

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

Appendix A

Appendix A.1. Transformer

The core of the Transformer is the scaled dot-product attention, which computes attention weights by projecting the input into a query matrix Q , key matrix K , and value matrix V , where the dimension of the keys is d_k . The attention mechanism is defined as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (\text{A1})$$

where the scaling factor $\sqrt{d_k}$ ensures gradient stability, and the softmax function normalizes the weights over values.

To enhance the representation capacity, the multi-head attention mechanism projects the input into h distinct subspaces, generating multiple sets of Q , K , and V . The scaled dot-product attention is computed independently for each head, and the results are concatenated and linearly transformed:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O, \quad (\text{A2})$$

with

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (\text{A3})$$

This mechanism enables the model to jointly capture diverse relational patterns across different subspaces, thereby enriching feature extraction.

Since the Transformer lacks inherent recurrence or convolution, positional encoding is introduced to incorporate the sequence order. Typically, sinusoidal functions are used to encode positional information and added to the input embeddings. In addition, each

encoder and decoder layer incorporates a position-wise FFN, applied independently to each position:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2, \quad (\text{A4})$$

where ReLU denotes the activation function. This FNN further refines the representations after attention computation.

Appendix A.2. ASSA

The structure of ASSA is given in Figure 2. Specifically, a squared ReLU-based self-attention mechanism (SSA) is introduced to filter out features with low query–key matching scores:

$$\text{Attention}_R = \text{ReLU}^2\left(\frac{QK^\top}{\sqrt{d_k}} + \text{Bias}\right), \quad (\text{A5})$$

where *Bias* denotes the relative position bias. In parallel, a dense self-attention branch (DSA), formulated similarly to Equation (A6), applies a softmax operation to preserve essential information and compensate for the potential over-sparsity of SSA:

$$\text{Attention}_S = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \text{Bias}\right). \quad (\text{A6})$$

The final ASSA attention output is obtained by adaptively fusing the two branches:

$$\text{Attention}_{\text{ASSA}} = (W_1 \cdot \text{Attention}_S + W_2 \cdot \text{Attention}_R)V, \quad (\text{A7})$$

where the normalized weights are defined as $W_1 = \frac{\exp(\alpha_1)}{\exp(\alpha_1) + \exp(\alpha_2)}$ and $W_2 = 1 - W_1$. Here, α_1 and α_2 are learnable parameters initialized with one of the two branches. This design provides a balanced mechanism that suppresses irrelevant noise interactions while exploiting sufficient informative features.

Appendix A.3. The FRFN

The partial convolution part of the FRFN can be expressed as

$$X_{\text{partial}} = \text{Conv1D}(X[:, :, \text{partial_channels}]), \quad (\text{A8})$$

where X represent the input of the FRFN. We use a one-dimensional CNN (Conv1D) for time series of PV data. The gating mechanism generates a gating signal through convolution and uses the Sigmoid activation function to limit the weight to the interval $[0, 1]$ to control the information flow of each time step or channel, given as

$$\text{Gate} = \text{Sigmoid}(\text{Conv1D}(x)), \quad (\text{A9})$$

and

$$X_{\text{out}} = X_{\text{partial}} \odot \text{Gate} + X, \quad (\text{A10})$$

which automatically suppresses irrelevant or noisy time steps and enhances the robustness of the model. Through this residual connection, the input can be passed directly to the output, alleviating the gradient vanishing problem and ensuring that the deep network can still preserve the long-term dependencies of the original time series.

References

1. Tina, G.M.; Ventura, C.; Ferlito, S.; De Vito, S. A state-of-art-review on machine-learning based methods for PV. *Appl. Sci.* **2021**, *11*, 7550. [CrossRef]
2. Mansouri, M.; Trabelsi, M.; Nounou, H.; Nounou, M. Deep learning-based fault diagnosis of photovoltaic systems: A comprehensive review and enhancement prospects. *IEEE Access* **2021**, *9*, 126286–126306. [CrossRef]
3. Yang, Y.L.; Che, J.X.; Deng, C.Z.; Li, L. Sequential grid approach based support vector regression for short-term electric load forecasting. *Appl. Energy* **2019**, *238*, 1010–1021. [CrossRef]
4. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy* **2018**, *164*, 465–474. [CrossRef]
5. Wang, J.; Li, P.; Ran, R.; Che, Y.; Zhou, Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. *Appl. Sci.* **2018**, *8*, 689. [CrossRef]
6. Ramkumar, G.; Sahoo, S.; Amirthalakshmi, T.; Ramesh, S.; Prabu, R.T.; Kasirajan, K.; Samrot, A.V.; Ranjith, A. A short-term solar photovoltaic power optimized prediction interval model based on FOS-ELM algorithm. *Int. J. Photoenergy* **2021**, 3981456. [CrossRef]
7. Chen, C.; Duan, S.; Cai, T.; Liu, B. Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Sol. Energy* **2011**, *85*, 2856–2870. [CrossRef]
8. Almonacid, F.; Pérez-Higueras, P.; Fernández, E.F.; Hontoria, L. A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator. *Energy Convers. Manag.* **2014**, *85*, 389–398. [CrossRef]
9. Vaz, A.; Elsinga, B.; Van Sark, W.; Brito, M. An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in Utrecht, the Netherlands. *Renew. Energy* **2016**, *85*, 631–641. [CrossRef]
10. Wang, K.; Qi, X.; Liu, H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. *Appl. Energy* **2019**, *251*, 113315. [CrossRef]
11. Jung, Y.; Jung, J.; Kim, B.; Han, S. Long short-term memory recurrent neural network for modeling temporal patterns in long-term power forecasting for solar PV facilities: Case study of South Korea. *J. Clean. Prod.* **2020**, *250*, 119476. [CrossRef]
12. Massaoudi, M.; Chihi, I.; Abu-Rub, H.; Refaat, S.S.; Oueslati, F.S. Convergence of photovoltaic power forecasting and deep learning: State-of-art review. *IEEE Access* **2021**, *9*, 136593–136615. [CrossRef]
13. Huang, X.; Li, Q.; Tai, Y.; Chen, Z.; Liu, J.; Shi, J.; Liu, W. Time series forecasting for hourly photovoltaic power using conditional generative adversarial network and Bi-LSTM. *Energy* **2022**, *246*, 123403. [CrossRef]
14. Yu, Y.; Cao, J.; Zhu, J. An LSTM short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access* **2019**, *7*, 145651–145666. [CrossRef]
15. Chen, Y.; Li, X.; Zhao, S. A Novel Photovoltaic Power Prediction Method Based on a Long Short-Term Memory Network Optimized by an Improved Sparrow Search Algorithm. *Electronics* **2024**, *13*, 993. [CrossRef]
16. Min, H.; Hong, S.; Song, J.; Son, B.; Noh, B.; Moon, J. SolarFlux Predictor: A Novel Deep Learning Approach for Photovoltaic Power Forecasting in South Korea. *Electronics* **2024**, *13*, 2071. [CrossRef]
17. Radhi, S.M.; Al-Majidi, S.D.; Abbod, M.F.; Al-Rawashidy, H.S. Machine Learning Approaches for Short-Term Photovoltaic Power Forecasting. *Energies* **2024**, *17*, 4301. [CrossRef]
18. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37. Available online: <https://ojs.aaai.org/index.php/AAAI/article/view/26317> (accessed on 8 October 2025).
19. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tao, D. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
20. Wu, J.; Zhao, Y.; Zhang, R.; Li, X.; Wu, Y. Application of three Transformer neural networks for short-term photovoltaic power prediction: A case study. *Sol. Compass* **2024**, *12*, 100089. [CrossRef]
21. Moon, J. A Multi-Step-Ahead Photovoltaic Power Forecasting Approach Using One-Dimensional Convolutional Neural Networks and Transformer. *Electronics* **2024**, *13*, 2007. [CrossRef]
22. Kim, J.; Obregon, J.; Park, H.; Jung, J.Y. Multi-step photovoltaic power forecasting using transformer and recurrent neural networks. *Renew. Sustain. Energ. Rev.* **2024**, *200*, 114479. [CrossRef]
23. Jing, S.; Xi, X.; Su, D.; Han, Z.; Wang, D. Spatio-Temporal Photovoltaic Power Prediction with Fourier Graph Neural Network. *Electronics* **2024**, *13*, 4988. [CrossRef]
24. Zhai, C.; He, X.; Cao, Z.; Abdou-Tankari, M.; Wang, Y.; Zhang, M. Photovoltaic power forecasting based on VMD-SSA-Transformer: Multidimensional analysis of dataset length, weather mutation and forecast accuracy. *Energy* **2025**, *324*, 135971. [CrossRef]
25. Xu, S.; Ma, H.; Ekanayake, C.; Cui, Y. Swin transformer-based transferable PV forecasting for new PV sites with insufficient PV generation data. *Renew. Energy* **2025**, *246*, 122824. [CrossRef]
26. Tang, H.; Kang, F.; Li, X.; Sun, Y. Short-term photovoltaic power prediction model based on feature construction and improved transformer. *Energy* **2025**, *320*, 135213. [CrossRef]

27. Liu, M.; Rao, S.; Huang, M.; Deng, M. Short-term photovoltaic power forecasting based on improved transformer with feature enhancement. *Sustain. Energy Grids Netw.* **2025**, *43*, 101759. [[CrossRef](#)]
28. Zhou, S.; Chen, D.; Pan, J.; Shi, J.; Yang, J. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2952–2963. [[CrossRef](#)]
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, ICLR, Vienna, Austria, 4 May 2021. Available online: <https://openreview.net/forum?id=YicbFdNTTy> (accessed on 8 October 2025).
30. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693. [[CrossRef](#)]
31. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, don't walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031. [[CrossRef](#)]
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735. [[CrossRef](#)]
33. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649. [[CrossRef](#)]
34. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555. [[CrossRef](#)]
35. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165. [[CrossRef](#)]
36. Zi, X.; Liu, F.; Liu, M.; Wang, Y. A Deep Learning Method for Photovoltaic Power Generation Forecasting Based on a Time-Series Dense Encoder. *Energies* **2025**, *18*, 2434. [[CrossRef](#)]

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.