



ARTICLE

Research on Ultra-Short-Term Photovoltaic Power Forecasting Method Based on Decision-Level Multimodal Fusion

Yong Xie¹, Shidong Feng¹, Wenzheng Yu^{2,*}, Yifeng Ma^{2,3}, Jianzhou Lin², Zhiyuan You⁴, Zhanyuan Ma¹, Fa Wen¹, Zhidan Hui¹ and Guoyu Niu⁵

¹China Resources New Energy (Qinghai) Co., Ltd., Xining, China

²School of Geographical Sciences, Nanjing University of Information Science and Technology, Nanjing, China

³Faculty of Engineering, University of Nottingham, University Park, Nottingham, UK

⁴Yangzhou Meteorological Bureau, Yangzhou, China

⁵China Resources Wind Power (Da Qaidam) Co., Ltd., Haixi, China

*Corresponding Author: Wenzheng Yu. Email: ywzheng519@126.com

Received: 06 August 2026; Accepted: 23 September 2026; Published: 18 October 2026

ABSTRACT: To address the problems of limited prediction accuracy of photovoltaic (PV) power forecasting under complex meteorological conditions when using a single data source or a single model, an ultra-short-term PV power forecasting method based on decision-level multimodal fusion is proposed. This method constructs two parallel prediction paths. Path 1 adopts a two-stage architecture of “first predicting cloud images, then regressing to power” to establish a cloud-image-to-power prediction model. Path 2 takes multi-source meteorological data as input and directly predicts PV power using the CatBoost model. At the decision level, the outputs of the two paths are combined through weighted fusion with fixed weights. Validated on measured data, the fusion model achieves a normalized root mean squared error (NRMSE) of 0.0608, a normalized mean absolute error (NMAE) of 0.0382, and a coefficient of determination (R^2) of 0.9550. All accuracy metrics are significantly better than those of any single-modality prediction model.

KEYWORDS: Photovoltaic power forecasting; ground-based cloud images; ultra-short-term forecasting; decision-level multimodal fusion; CatBoost; ResSE-STNet; channel attention mechanism

1 Introduction

In recent years, the rapid development of renewable energy sources such as photovoltaic (PV) power generation has effectively alleviated the energy crisis. However, the randomness and volatility of PV output pose challenges to the stable operation of the power grid [1,2]. The stochasticity and volatility of photovoltaic (PV) power output are rooted in underlying physical mechanisms: the output power emerges from the multivariate coupling among irradiance, module temperature, and system operating conditions. Serving as the primary energy input, irradiance dictates the power amplitude and exhibits rapid, second-to-minute scale fluctuations driven by cloud movement and meteorological variations. Meanwhile, module temperature—jointly determined by ambient temperature and irradiance-induced heating—exhibits a minute-level thermal inertia. When superimposed with the modulating effects of operating conditions such as Maximum Power Point Tracking (MPPT) deviations, partial shading, and active power curtailment, these elements collectively shape the nonlinear and multi-scale fluctuation characteristics of the PV output, thereby exacerbating the

difficulty of power forecasting [3]. Accurate PV power forecasting can help dispatching departments adjust their plans and mitigate the negative impacts of grid-connected PV systems [4–6].

A substantial amount of research has been conducted by scholars worldwide on solar power forecasting, leading to three main categories of methods: physical models, traditional statistical methods, and machine learning methods. Physical models typically rely on detailed physical parameters [7], such as the structure and installation angle of PV modules. However, due to the difficulty of obtaining these parameters, statistical models have gradually become the mainstream approach. Statistical models extract the fluctuation characteristics of PV power output from historical data and establish mathematical models, providing high prediction accuracy without requiring detailed physical information. Consequently, they have been widely applied in PV power forecasting. For example, Babatunde et al. employed multiple linear regression models to handle correlation indicators [8]; Ding Ming et al. constructed a forecasting model based on Markov chains [9]; He Weiguo et al. combined SVR with the ARIMA model to achieve ultra-short-term forecasting [10]; other statistical approaches such as grey system models and moving average sequence models have also been utilized [11–14]. Although these statistical methods have improved prediction accuracy, they often face challenges such as difficulty in modeling complex nonlinear relationships and poor capability in fusing multimodal heterogeneous data, which limit their broader application. In contrast, machine learning methods, especially deep learning algorithms, have demonstrated significant advantages in the field of solar power forecasting. Malvoni et al. [15] combined PCA-based dimensionality reduction with a least-squares SVM (LSSVM) to enhance prediction performance; Xiang Mingjun [16] integrated VAE and LSTM with meteorological and cloud image data to forecast power; Zhan Xu et al. attempted to improve satellite cloud image prediction using GANs [17]; and Shree et al. developed an LSTM-based model that detects similar days according to the influence of different weather variables, thereby forecasting PV power under different weather conditions [18]. More recently, deep spatiotemporal architectures have been extensively explored for image-driven ultra-short-term forecasting: Zang et al. proposed a two-stream network with a PV-guided attention mechanism to explicitly model the interaction between the spatial and temporal features extracted from ground-based sky images [19]; Xu et al. applied a vision transformer to sky images for ultra-short-term PV generation forecasting [20]; and Huang et al. developed a 3D ConvLSTM-CNN model that exploits multi-channel colour information from consecutive all-sky images for irradiance forecasting [21]. In parallel, hybrid deep learning frameworks that couple signal decomposition with Transformer-based temporal modelling have also been reported for PV power forecasting [22].

Although effective PV forecasting methods have been developed, using a single prediction model under variable weather conditions may lead to larger errors, especially on cloudy or rainy days. To address these challenges, this paper proposes a ultra-short-term forecasting model that integrates CatBoost with ResSE-STNet-ResNet fusion. The model designs two parallel paths to optimize the accuracy of PV power forecasting.

The first path constructs a mapping model based on ground-based cloud images and PV power generation data, achieving refined PV output prediction by analyzing visual features such as cloud cover and movement trends. This design follows the indirect, “predict-then-regress” paradigm also adopted by recent generative approaches, in which future sky images are first synthesized from historical sequences and then mapped to PV output [23]; its relative merits with respect to direct end-to-end image-to-power regression have been systematically benchmarked in the recent literature [24]. The second path directly establishes a mapping relationship between multi-source meteorological data (including irradiance, temperature, wind speed, humidity, etc.) and PV power generation, capturing the direct impact of environmental factors on power generation efficiency. After each path independently generates PV power prediction results, the

decision level feeds these results into an integration module, which outputs the final PV power mapping through a weighted fusion mechanism. This design not only effectively avoids the limitations of a single model but also significantly enhances prediction stability under a variety of weather conditions, providing more reliable data support for grid dispatching and energy management. Finally, a case study is conducted using actual PV data to verify the effectiveness of the proposed method.

2 Methodology

2.1 SE Attention Mechanism

The SE attention mechanism is a channel-adaptive feature enhancement method that optimizes feature representation by dynamically learning importance weights across channels [25]. Its core comprises two stages: in the Squeeze stage, global average pooling compresses the spatial dimensions, aggregating the global spatial information of each channel into a scalar; in the Excitation stage, two fully connected layers are employed to construct nonlinear interaction relationships among channels—first reducing dimensionality then expanding it—and a Sigmoid function generates channel weight vectors ranging from 0 to 1, which are finally used to perform channel-wise weighting of the original features. This mechanism can autonomously strengthen key feature channels relevant to the task (e.g., cloud movement regions) while suppressing redundant or noisy channels, thereby significantly enhancing feature discriminability within lightweight models. It is especially suitable for dynamically focusing on meteorologically sensitive regions in cloud image forecasting tasks. The specific formula is as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j) \quad (1)$$

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \quad (2)$$

$$\tilde{x}_c = s_c \cdot x_c \quad (3)$$

where z_c denotes the compressed value of the c -th channel; W_1 is the dimensionality reduction matrix; W_2 is the dimensionality expansion matrix; δ is the ReLU activation function; σ is the Sigmoid function; $\tilde{x}_c$ is the learned feature; and s_c is the importance weight of the channel, with values closer to 1 indicating greater importance.

2.2 Residual Block

In machine learning, the residual refers to the difference or error between the model's predicted value for a sample and the true observed value of that sample. It represents the portion of information that the model fails to fully explain or capture at the current data point, which may arise from data noise, complex patterns that the model has not learned, omitted important features, or inherent limitations of the model. By implementing identity mapping through cross-layer skip connections, the network's learning objective is transformed into a residual function that is easier to optimize, thereby effectively alleviating the vanishing/exploding gradient problem, avoiding the degradation of deep networks, significantly improving training stability and model expressiveness, and making it possible to construct high-performance extremely deep networks [26]. The formula for the residual is as follows:

$$f(x) = x + g(x) \quad (4)$$

where $f(x)$ denotes the output of the residual block; x denotes the input (identity); and $g(x)$ denotes the residual mapping.

2.3 Spatiotemporal Mechanism

In this model, the spatiotemporal module achieves comprehensive modeling of both spatial structure and temporal variations in image sequences through a three-stage architecture comprising a spatial feature extractor, a spatiotemporal fusion module, and a motion decoder.

First, the spatial feature extractor independently applies a series of convolution, normalization, and residual connection operations to each input frame, extracting multi-scale, multi-channel local texture and geometric information layer by layer. Meanwhile, channel attention is introduced at this stage to adaptively weight important feature channels, further enhancing the network's sensitivity to fine-grained structures.

Subsequently, the spatiotemporal fusion module concatenates multi-frame spatial features along the channel dimension and performs cross-frame information aggregation on the concatenated features using large convolutional kernels. Attention mechanisms and residual structures are again incorporated, enabling temporal context (such as object motion direction and velocity changes) to be fully interacted with and redistributed at the channel level.

Finally, the motion decoder employs deconvolution and convolution layers to restore spatial resolution, progressively reconstructing a pixel-level prediction of the next frame. Immediately after, a frame generator maps the decoded features into a three-channel image space through small convolutional kernels and a smooth activation function, outputting a predicted frame that conforms to the statistical distribution of real scenes.

Overall, this spatiotemporal module strikes a balance between extracting multi-scale spatial structural features and enabling efficient interaction of inter-frame dynamic information through concatenation and attention-based fusion. As a result, it accurately captures the motion evolution patterns across consecutive frames while preserving spatial details.

2.4 ResSE-STNet Model

ResSE-STNet is a deep learning model specifically designed for ground-based cloud image sequence prediction, with the core objective of accurately inferring future cloud image distributions from consecutive historical frames. The model adopts an end-to-end training architecture. The overall pipeline is shown in Fig. 1, and consists of four main stages in sequence: a spatial feature extraction module, a spatiotemporal fusion module, a motion decoder, and a frame generator. The model takes six consecutive historical ground-based cloud image frames as input, and through the collaborative operation of the four core modules, outputs the predicted frame for the next time step.

Fig. 2 illustrates the individual modules of the ResSE-STNet model. In the spatial feature extraction stage, the model processes each frame of the input sequence independently. This module is built upon a residual structure and gradually extracts multi-scale spatial features through convolutional layers, normalization layers, and skip connections. Notably, a channel attention mechanism is embedded in the module, which adaptively calibrates the weights of each feature channel through squeeze-and-excitation operations, thereby enhancing feature responses sensitive to cloud motion while suppressing redundant information. After two downsampling operations, the spatial resolution of the feature maps is reduced to 1/4 of the input.

The spatiotemporal fusion module is responsible for aggregating cross-frame temporal information. This module concatenates the previously extracted spatial features of the six frames along the channel dimension, forming a joint representation that fuses multi-frame information. Subsequently, large-kernel

convolution operations are employed to capture inter-frame dependencies, and attention mechanisms along with residual connections are again introduced to model the motion trends and evolution patterns of cloud systems.

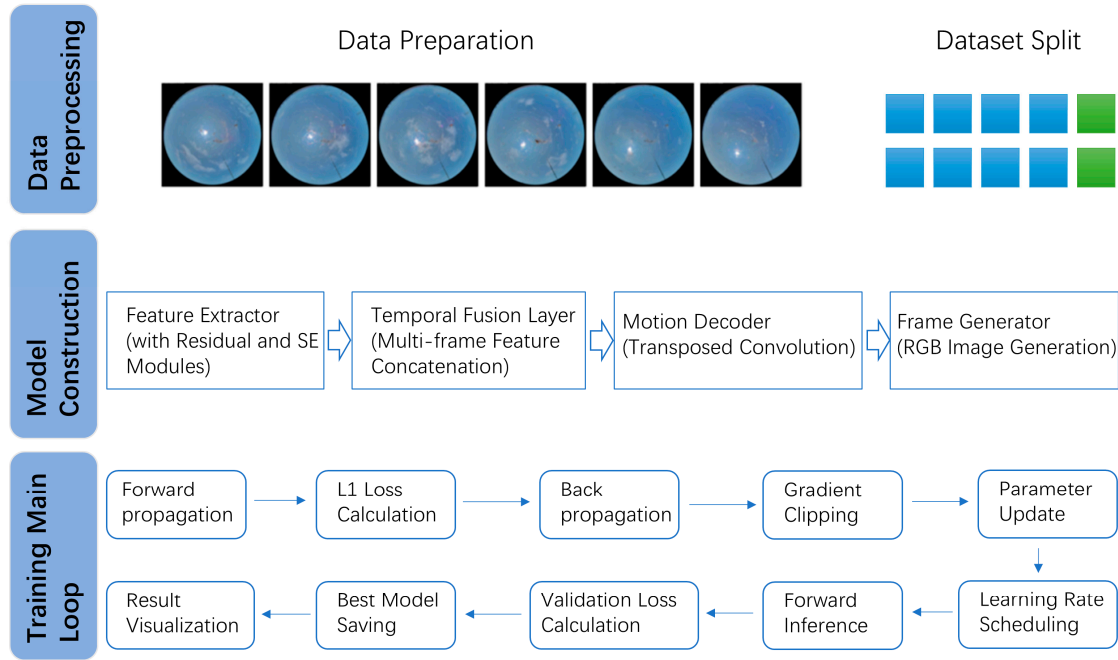


Figure 1: Prediction pipeline of the ResSE-STNet model.

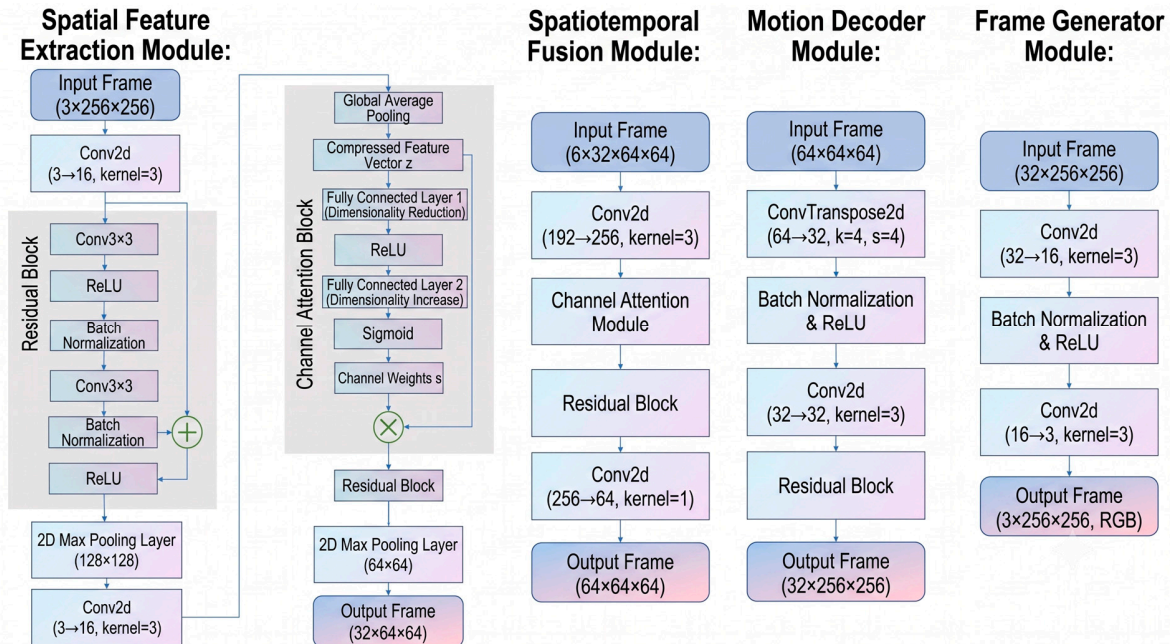


Figure 2: Modules of the ResSE-STNet model.

The motion decoder undertakes the tasks of feature map upsampling and spatial detail reconstruction. This module achieves $4\times$ upsampling through transposed convolution, gradually restoring the spatial resolution of the feature maps, while incorporating residual structures to preserve deep semantic information.

Finally, the frame generator maps the high-dimensional features into a three-channel RGB image via a convolutional layer and uses a Sigmoid activation function to constrain pixel values within the range of $[0, 1]$, thereby producing the final visualized prediction frame.

With regard to model training, the L1 loss function is employed to measure the pixel-level difference between the predicted frame and the ground truth frame, supplemented by the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) as auxiliary metrics for evaluating the visual quality of the predicted images. The optimization process utilizes the AdamW optimizer [27,28] and adopts a cosine annealing schedule to dynamically adjust the learning rate [29], facilitating model convergence. To enhance training efficiency and stability, mixed precision training and gradient clipping techniques are introduced [30,31], and an early stopping mechanism is implemented to prevent overfitting.

In this study, a mapping model based on deep residual networks is constructed to achieve end-to-end regression modeling from ground-based cloud images to PV power. The model backbone adopts ResNet-18 to capture general patterns such as cloud structure, texture, and illumination distribution. The input layer receives normalized cloud images at a resolution of 256×256 . After initial convolution and pooling, hierarchical feature extraction is performed through four stages of residual modules, where each module mitigates the gradient degradation problem in deep networks via residual connections. The regression head is restructured into a dual-hidden-layer architecture: the original classification layer is replaced with a mapping from 512 dimensions to 128 dimensions, incorporating ReLU activation and a dropout rate of 0.3 for regularization, and finally the output layer is compressed to a single power prediction value. The training process employs the Adam optimizer with a fixed learning rate of $1e-4$ and a batch size of 32. The dataset is divided into training, validation, and test subsets in a ratio of 7:1.5:1.5. To avoid overfitting, an early stopping strategy is implemented, terminating training if the validation loss does not improve for 50 consecutive epochs. The model is optimized with the mean squared error loss and comprehensively evaluated on the test set using metrics such as MAE, RMSE, and R^2 . Through discriminative representation learning, this design automatically models the nonlinear coupling between high-level semantics—such as cloud cover, cloud pattern evolution, and the relative position of the sun—and power fluctuations. Compared with traditional manual feature engineering, it demonstrates stronger dynamic modeling capability and generalization performance in rapidly changing cloud scenarios.

2.5 ResNet Model

Residual Network (ResNet) is a revolutionary deep learning architecture that addresses the vanishing gradient and network degradation problems in deep networks by introducing residual learning and skip connections, Fig. 3 presents a schematic diagram of the residual network. Its core idea is to have the network learn the residual between the input and the target, rather than a direct mapping, and to preserve original feature information through cross-layer identity connections, making the training of networks with thousands of layers feasible. A typical residual block consists of convolution, batch normalization, and ReLU activation. This structure has achieved breakthrough performance in tasks such as image classification and object detection and has become a cornerstone of modern deep models, with its residual bottleneck module striking a balance between computational efficiency and accuracy. Its calculation formula is shown in Eq. (4).

In this study, a mapping model based on deep residual networks is constructed to achieve end-to-end regression modeling from ground-based cloud images to PV power. The model backbone adopts ResNet-18 to capture general patterns such as cloud structure, texture, and illumination distribution. The input layer receives normalized cloud images at a resolution of 256×256 . After initial convolution and pooling,

hierarchical feature extraction is performed through four stages of residual modules, where each module mitigates the gradient degradation problem in deep networks via residual connections. The regression head is restructured into a dual-hidden-layer architecture: the original classification layer is replaced with a mapping from 512 dimensions to 128 dimensions, incorporating ReLU activation and a dropout rate of 0.3 for regularization, and finally the output layer is compressed to a single power prediction value. The training process employs the Adam optimizer with a fixed learning rate of $1e-4$ and a batch size of 32. The dataset is divided into training, validation, and test subsets in a ratio of 7:1.5:1.5. To avoid overfitting, an early stopping strategy is implemented, terminating training if the validation loss does not improve for 50 consecutive epochs. The model is optimized with the mean squared error loss and comprehensively evaluated on the test set using metrics such as MAE, RMSE, and R^2 . Through discriminative representation learning, this design automatically models the nonlinear coupling between high-level semantics—such as cloud cover, cloud pattern evolution, and the relative position of the sun—and power fluctuations. Compared with traditional manual feature engineering, it demonstrates stronger dynamic modeling capability and generalization performance in rapidly changing cloud scenarios.

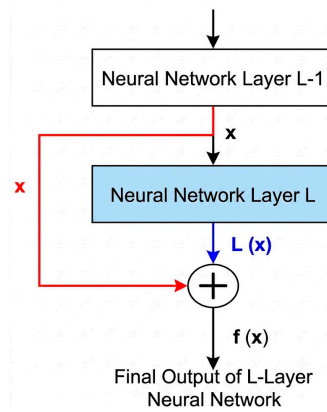


Figure 3: Schematic diagram of the residual network.

2.6 Tree Models

Tree models, as a widely used class of algorithms in the field of machine learning, accomplish prediction or classification tasks by recursively partitioning the feature space [32]. Their core advantages are multifaceted. First, the models possess a high degree of interpretability [33]: the decision-making process can be intuitively transformed into tree-shaped rules, facilitating understanding and trust, which makes them particularly valuable in scenarios such as financial risk control and medical diagnosis where model transparency is strictly required. Second, tree models have low dependence on data preprocessing; they do not require rigorous standardization, can naturally handle both numerical and categorical features, and exhibit strong robustness to missing values and outliers. In terms of pattern learning, tree models can automatically capture interactions among features and effectively fit nonlinear relationships, which is especially prominent when the intrinsic regularities of the data exhibit complex nonlinear structures. Furthermore, by integrating multiple decision trees through ensemble learning frameworks (e.g., Random Forest [34], Gradient Boosting Decision Trees such as GBDT/XGBoost, etc.), tree models significantly improve both the accuracy and stability of predictive performance while effectively alleviating the inherent overfitting tendency of a single decision tree. These characteristics collectively establish the important role of tree models in practical applications.

This study employs the CatBoost model, which possesses outstanding capability for handling structured tabular data. As an ensemble tree model, CatBoost achieves an optimal balance between accuracy and efficiency by virtue of its powerful feature fitting capacity and excellent computational efficiency [35].

3 Case Study

3.1 Dataset Description

This study takes the Sheyang 2# photovoltaic (PV) power station (with an installed capacity of 100 MW, central coordinates 33.18° N, 119.59° E) located in Baoying County, Yangzhou City, Jiangsu Province, as the research site. The data collection period spans from 1 May to 29 July 2020, during which surface irradiance data, ground-based sky images, and actual PV power output data were continuously collected for three months. The PV power data were sampled at 15-min intervals. In addition, this study also utilizes reanalysis meteorological data from the National Solar Radiation Database (NSRDB) at the same location (spatial resolution of 0.01°) [36]. This database updates meteorological and radiation parameters every 10 min, covering global horizontal irradiance, diffuse horizontal irradiance, direct normal irradiance, wind speed, surface pressure, relative humidity, and temperature. The meteorological data used in this study are shown in Table 1. A ground-based all-sky imager installed on-site captures a sky image every 10 min from 6:00 to 19:50 daily, providing abundant information on cloud cover variations for the research. To eliminate discrepancies in temporal resolution between the PV power and meteorological data, the time-series data were resampled to a 30-min resolution and chronologically partitioned into training, validation, and testing sets, comprising 70%, 15%, and 15% of the dataset, respectively. The forecasting task is defined as a single-step deterministic rolling prediction model with a 30-min prediction horizon. To ensure reproducibility and exclude information leakage, all modalities are aligned to a common 30-min grid. Let t_0 be the forecast issue time and $t_0 + 30$ min the target time. The cloud-image path uses the six most recent resampled frames up to t_0 . The CatBoost path uses meteorological variables, Power Station Irradiance (PSI), and historical power observed up to t_0 . PSI is a measured value at the PV plant and is available at t_0 . No input later than t_0 is used. The dataset is split chronologically into training, validation, and test sets. All data-driven preprocessing parameters, including the MLP imputation model, K-means clustering centers, and normalization ranges, were fitted exclusively on the training set and then applied unchanged to the validation and test sets.

Table 1: Specifications of meteorological feature data.

NAME	ABBR.	UNIT
Global Horizontal Irradiance	GHI	W/m ²
Diffuse Horizontal Irradiance	DHI	W/m ²
Direct Normal Irradiance	DNI	W/m ²
Clear Sky Global Horizontal Irradiance	CGHI	W/m ²
Clear Sky Global Horizontal Irradiance	CDHI	W/m ²
Clear Sky Direct Normal Irradiance	CDNI	W/m ²
Temperature	TEMP	°C
Dew Point Temperature	DEWP	°C
Wind Speed	WS	m/s
Atmospheric Pressure	P	mbar
Precipitable Water	PW	cm
Relative Humidity	RH	%

3.2 Dataset Preprocessing

3.2.1 Ground-Based Cloud Image Processing

Cloud images acquired by ground-based all-sky imagers serve as a critical foundational data source for research and applications such as cloud cover estimation, cloud classification, and solar radiation forecasting. However, raw cloud image data often suffer from missing-value issues, including pixel-level missing and complete frame loss, as well as various occlusions caused by lens frames, device supports, birds, dirt, and even overexposed solar glare. Therefore, efficient and precise data preprocessing—particularly imputation and reconstruction for missing values and the identification and removal of various occlusions—is a primary step to ensure the accuracy and reliability of subsequent analyses.

In this study, since the number of missing cloud images is relatively small and the cloud image data exhibit temporal continuity, we employ an optical flow-based method to process the missing images by utilizing the cloud images acquired 10 min before and 10 min after each missing frame. The core principle of generating an intermediate-time image via optical flow is as follows: first, a dense optical flow field (i.e., pixel-level motion vectors) is estimated from the two adjacent frames under the assumption of constant-velocity motion over a short time interval. Subsequently, the optical flow vectors are decomposed proportionally to the time interval (e.g., a 50% displacement is used when generating the middle frame). A bidirectional projection fusion is then performed: pixels from the previous frame are projected to the intermediate position along the forward-decomposed flow displacement, and simultaneously pixels from the subsequent frame are projected to the same position along the backward-decomposed flow displacement. Finally, a weighted average of the two projection results is computed (with dynamically adjustable weights for occluded regions, for instance), supplemented by interpolation techniques to fill any remaining gaps, thereby synthesizing the target image. The interpolation results are illustrated in Fig. 4.

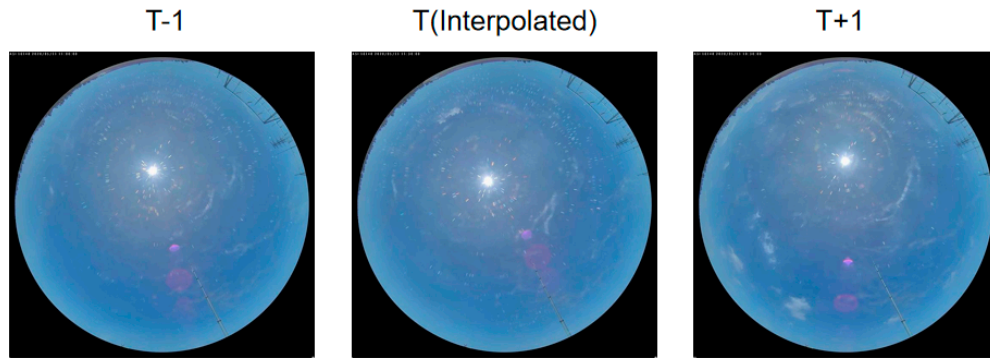


Figure 4: Results of image interpolation using the optical flow method.

In this study, missing cloud images occurred only in the training set. They were reconstructed using an optical-flow-based method applied within the training set only. The optical-flow method is parameter-free and involves no data-driven fitting. The test set cloud images are complete, and no imputation or reconstruction was performed on them. Therefore, no validation or test observations were used in this preprocessing step.

3.2.2 Preprocessing of Meteorological and Photovoltaic Power Generation Data

Meteorological data play a critical role in influencing photovoltaic power generation, encompassing multiple environmental parameters such as temperature, humidity, wind speed, and solar irradiance. The dataset adopted in this study includes not only the core radiative quantities that directly determine the

energy input to photovoltaic systems—namely global horizontal irradiance, diffuse horizontal irradiance, and direct normal irradiance—but also key environmental parameters including air temperature, dew point temperature, wind speed, atmospheric pressure, precipitable water vapor, and relative humidity. For missing values in the meteorological power data, several machine learning methods were employed for imputation, and the multilayer perceptron (MLP) was ultimately selected to fit the missing data. The K-means clustering algorithm was used to detect outliers; once identified, outliers were treated as missing values and subsequently imputed using the MLP-based missing-data handling method. The MLP imputation model and the K-means anomaly detector were trained only on the training set. Outliers identified in the validation and test sets were treated as missing values and imputed using the training-fitted MLP. No validation or test information was used to fit these preprocessing models.

3.2.3 Data Normalization

In data preprocessing, normalization methods such as z-score standardization and 0–1 scaling are commonly employed to eliminate dimensional discrepancies. The former is suitable for normally distributed data, while the latter preserves the original relative relationships. Given that the focus of this study is photovoltaic power data, which are inherently positive and typically free of negative values or extreme outliers, the 0–1 normalization method is considered more appropriate. This approach not only effectively avoids biases induced by extreme values but also ensures that all photovoltaic power data are reasonably mapped onto a common scale, thereby improving the accuracy and stability of subsequent analyses or model training. Therefore, the 0–1 normalization method was adopted. The minimum and maximum values used for normalization were computed from the training set only, and the same transformation was then applied to the validation and test sets, avoiding any leakage of their value ranges.

3.3 Evaluation Metrics for Prediction

The performance of a predictive model can be evaluated using three metrics: the coefficient of determination, the normalized mean squared error, and the normalized root mean squared error. The coefficient of determination reflects the accuracy of model predictions, with values closer to 1 indicating more accurate predictions. The normalized mean squared error and the normalized root mean squared error eliminate dimensional effects and provide a standardized error metric, facilitating comparisons across different models. Together, these metrics constitute a framework for evaluating the performance of predictive models. Their specific formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$NMSE = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{(y_{\max} - y_{\min})^2} \quad (6)$$

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \quad (7)$$

$$NMAE = \frac{\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|}{y_{\max} - y_{\min}} \quad (8)$$

where n is the number of observations; y_i is the true value of the i -th observation; $\hat{y}_i$ is the predicted value corresponding to y_i ; $\bar{y}_i$ is the mean of all observed values; $y_{\max}$ is the maximum value among all observations; and $y_{\min}$ is the minimum value among all observations.

3.4 Evaluation of Ground-Based Cloud Image Generation

For image/video quality assessment, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are adopted as evaluation metrics. PSNR is primarily used to measure the degree of distortion between the original signal and the processed signal (e.g., after compression, denoising, or reconstruction) [37]. A PSNR greater than 30 dB generally indicates good compression quality, whereas a value below 20 dB may correspond to noticeable distortion visible to the naked eye. SSIM comprehensively evaluates the visual consistency between two images from three dimensions—luminance, contrast, and structure—by simulating the human visual perception mechanism of image content [38].

As shown in Fig. 5, the synergistic effect of the SE attention mechanism and the residual module significantly improves the prediction accuracy of ground-based cloud images. The performance metrics (Table 2) show that the ResSE-STNet model achieves the optimal values for both PSNR (30.34) and SSIM (0.9336), representing an approximately 13% improvement over the traditional optical flow method (PSNR 26.93), and outperforms single-module or baseline models. Specifically, the SE module strengthens key motion features such as cloud edge diffusion direction and texture density variations by dynamically adjusting channel weights, resulting in clearer cloud boundaries in the predicted frames and accurate detail restoration, particularly in high-dynamic regions of the input frame sequence (e.g., color discontinuities at cloud intersections). Meanwhile, the residual module effectively fuses shallow basic structures (e.g., the overall contour of cloud images) with deep temporal dynamics (e.g., cloud movement speed) through cross-layer skip connections, avoiding blurring caused by feature degradation in models without residual connections (e.g., cloud fragmentation in predicted frames). The combination of the two not only ensures stable training of deep networks but also achieves the collaborative optimization of feature selection and information transmission, ultimately attaining higher fidelity in cloud morphology details and motion trajectory prediction.

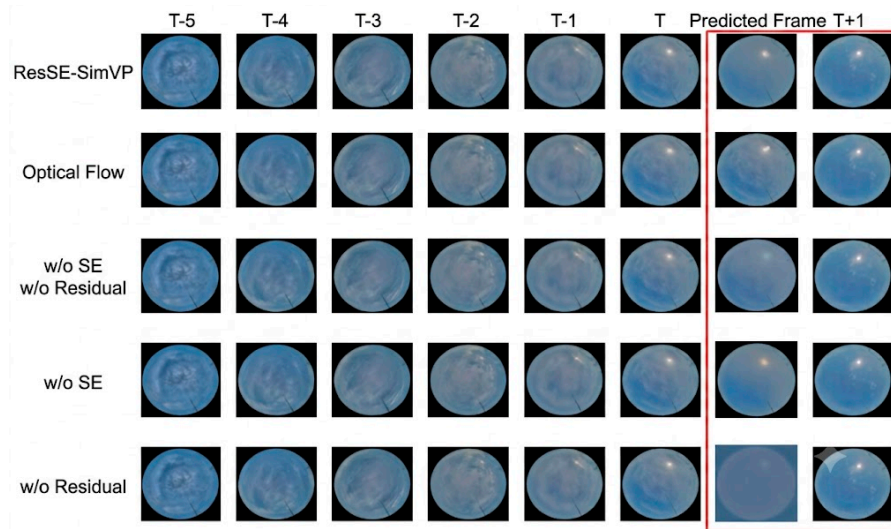


Figure 5: Comparison of ground-based cloud image predictions.

Table 2: Evaluation of prediction results.

Model	PSNR	SSIM
ResSE-STNet	30.34	0.9336
Without SE & without residual	30.06	0.9312
Without SE	29.51	0.9277
Without residual	30.12	0.9315
Optical flow method	26.93	0.9163

3.5 Evaluation of Regression from Ground-Based Cloud Images to PV Power

The model performing PV power regression based on ground-based cloud images using the ResNet network demonstrates excellent performance, Table 3 presents the error values of the model regression. Its normalized root mean squared error (NRMSE) is 0.0558 and normalized mean absolute error (NMAE) is 0.0356, indicating that the deviations between predicted and actual values are very small. Meanwhile, the model achieves a high coefficient of determination (R^2 score) of 0.9652, meaning that it can explain the vast majority of the power variation, exhibiting outstanding fitting performance. This verifies the strong correlation between the visual features of cloud images and the synchronized PV power.

Table 3: Regression error values.

Error Metric	Value
NMAE	0.0356
NMSE	0.0031
NRMSE	0.0558
R^2 Score	0.9652

3.6 Evaluation of Photovoltaic Power Forecasting Based on Ground-Based Cloud Images

The PV power forecasting method based on ground-based cloud images proposed in this study consists of two core stages that are sequentially connected. The first stage is cloud image sequence prediction, which employs the ResSE-STNet model. This model extracts spatial features of each cloud image frame through residual blocks embedded with the SE attention mechanism, and utilizes a spatiotemporal fusion module to capture dynamic evolution across consecutive frames, thereby generating high-quality forecasted ground-based cloud images at a specified future moment. The second stage is power regression prediction, in which the predicted cloud image output from the first stage is fed as input into a deep regression neural network based on the ResNet architecture. Through its hierarchical convolutional and residual structures, this network automatically learns the complex nonlinear mapping relationship between visual features such as cloud morphology and distribution in the predicted cloud image and PV power, ultimately outputting the corresponding point prediction of power. Table 4 presents the power prediction errors of the ResSE-STNet-ResNet model. Specifically, for the ground-based cloud image forecasting model, the input sequence length is 6 steps; for the power regression model based on the forecast cloud images, the input sequence length is 1 step. Fig. 6 presents the overall framework of the sequential two-stage PV power forecasting model. The architecture is fundamentally composed of two sequentially linked core phases: cloud image sequence prediction and power regression estimation.

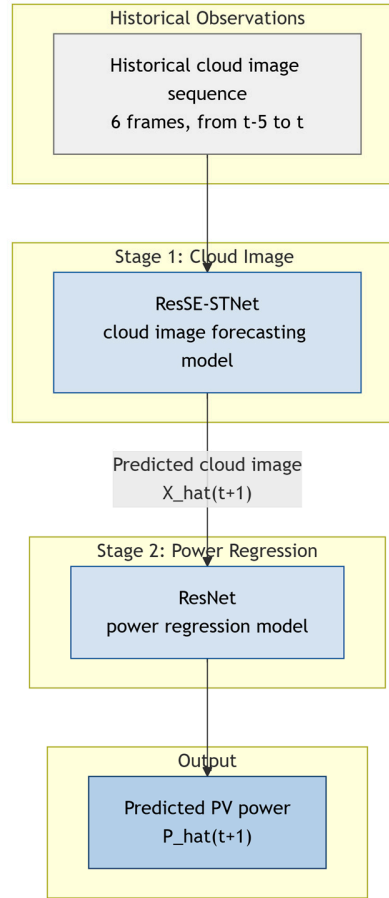


Figure 6: Schematic diagram of the sequential two-stage PV power forecasting model.

Table 4: Power prediction errors of the ResSE-STNet-ResNet model.

Error Metric	Value
NMAE	0.1198
NMSE	0.0308
NRMSE	0.1756
R ² Score	0.8605

3.7 Photovoltaic Power Forecasting Based on Meteorological Data

In meteorological data-driven photovoltaic power forecasting, total solar irradiance acts as the most direct driving factor, primarily determining the photo-generated current and output power of the PV modules. Specifically, the ratio of direct to diffuse irradiance dictates the effective irradiance actually received by the modules and their incidence angle response, while short-term cloud movement can induce severe minute-level power fluctuations. Ambient and module temperatures affect the open-circuit voltage and fill factor by altering the properties of the semiconductor materials; typically, an elevation in module temperature decreases the conversion efficiency according to a negative temperature coefficient. Conversely, wind speed lowers the module's operating temperature by enhancing convective heat transfer with the surrounding air, thereby mitigating thermal derating losses to some extent. As a result, wind gusts can bring about rapid, albeit minor, power variations on short time scales. Humidity influences the effective incident radiation and spectral transmittance primarily by altering atmospheric water vapor content, aerosol optical

properties, and the state of condensation or soiling on the module surface. Therefore, forecasting models must simultaneously account for the real-time values and rates of change of various meteorological variables, as well as their nonlinear coupling relationships with the electrical characteristics of the PV system.

In this stage, a PV power regression forecasting model based on the CatBoost algorithm is constructed. The hyperparameters of the CatBoost model are presented in Table 5. At the forecast issue time t_0 , the CatBoost model uses only inputs available at or before t_0 : meteorological variables, PSI and historical power. The target is the PV power at $t_0 + 30$ min. All inputs are aligned to t_0 on the common 30-min grid. The input sequence length is 1 step, i.e., the feature vector at t_0 . No future information is used. This setup aims to explore the complex nonlinear mapping relationship between meteorological conditions and PV output.

Table 5: CatBoost hyperparameters.

Hyperparameter	Value
iterations	1000
learning_rate	0.1
depth	6
loss_function	RMSE
eval_metric	RMSE
early_stopping_rounds	50
verbose	100
random_seed	42
use_best_model	True

A systematic data processing pipeline is adopted, performing missing value imputation and automatic feature type identification on the raw meteorological data. During the model construction phase, the dataset is split into training, validation, and test sets in a ratio of 70%-15%-15%, and the CatBoost regressor is configured with root mean squared error (RMSE) as both the optimization objective and the early stopping monitoring metric, effectively preventing model overfitting. The final metrics are shown in Table 6, where the normalized root mean squared error (NRMSE) is 0.0662 and the normalized mean absolute error (NMAE) is 0.0434.

Table 6: Error analysis.

Error Metric	Value
NMAE	0.0434
NMSE	0.0044
NRMSE	0.0662
R ² Score	0.9486

The feature importance analysis provides a strong physical explanation for the model's prediction logic, enhancing the credibility of the results. As shown in Fig. 7, Power Station Irradiance (PSI) emerges as the most critical predictor with a markedly higher importance value (approximately 14) compared to other features. Solar irradiance, encompassing both direct and diffuse components, serves as the fundamental driver of the photovoltaic effect and directly determines the short-circuit current of the modules. Consequently, abrupt fluctuations in irradiance induced by cloud movement constitute the primary cause of severe instantaneous volatility in PV power output. Following closely, historical power (Power) and Global Horizontal Irradiance (GHI) represent the system's inertial characteristics and real-time total radiation conditions, respectively, further consolidating the central role of radiation-related driving factors. In

contrast, features such as Diffuse Horizontal Irradiance (DHI), Temperature (TEMP), and Relative Humidity (RH) have relatively low importance scores. As secondary meteorological variables affecting photovoltaic generation, these variables do not dictate instantaneous, drastic power fluctuations; rather, they critically fine-tune the system's efficiency via optical and thermodynamic mechanisms. Acting as a supplement to direct sunlight, DHI provides the baseline photo-generated current during cloud cover, sustaining the system's generation baseline under low-light conditions. TEMP primarily exerts a "thermal derating" effect, wherein elevated temperatures decrease the module's open-circuit voltage, thereby imposing a gradual yet persistent constraint on photoelectric conversion efficiency. Meanwhile, RH intensifies light scattering via atmospheric moisture and can alter panel transmittance, resulting in minor extinction losses of the effective radiation reaching the modules. Although these three factors collectively dictate the operational upper limit of PV systems in complex real-world environments, this indicates that under the current data and modeling framework, they have relatively limited direct influence on instantaneous power fluctuations.

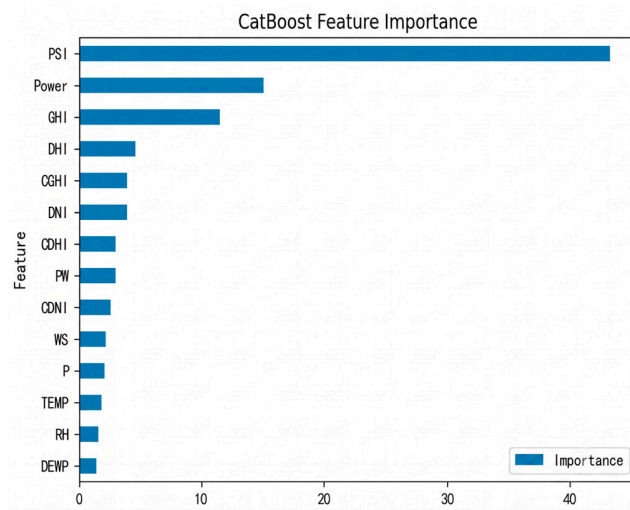


Figure 7: Model feature importances.

3.8 Photovoltaic Power Forecasting Based on Decision-Level Multimodal Fusion

Multimodal data fusion mainly includes three approaches: data-level fusion, feature-level fusion, and decision-level fusion. Among them, decision-level fusion models each modality independently and combines the prediction results at the decision stage through methods such as weighted voting, probability averaging, or ensemble learning [39]. Recent multimodal frameworks have likewise coupled predicted sky-image sequences with historical power series through Transformer-based fusion, in order to exploit the complementarity of the two modalities [40].

To integrate the advantages of the two independent prediction paths—ground-based cloud images and meteorological data—this section constructs a multimodal fusion forecasting method based on the decision level. This method does not mix low-level features but instead performs decision integration after each path generates its own power point forecasts. Both paths share the same forecast issue time t_0 , i.e., both use only inputs available at or before t_0 . No future information is used. Specifically, the power sequence output by the cloud image prediction model based on ResSE-STNet-ResNet and the power sequence output by the meteorological prediction model based on CatBoost are fed into a weighted fusion module, which produces the final forecast as a linear combination of the two path outputs:

$$\hat{P} = w_{\text{met}}\hat{P}_{\text{met}} + w_{\text{cloud}}\hat{P}_{\text{cloud}}, w_{\text{met}} + w_{\text{cloud}} = 1 \quad (9)$$

The two fusion weights are determined by an inverse-error (reciprocal-error) weighting rule rather than assigned empirically. Because a smaller value indicates better performance for all three adopted metrics, NMAE, NMSE and NRMSE are used directly as the error e of a path. For each metric m , the weight of path i is the reciprocal of its error, normalized over the two paths:

$$w_i^{(m)} = \frac{1/e_i^{(m)}}{\sum_{j \in \{\text{met}, \text{cloud}\}} 1/e_j^{(m)}}, i \in \{\text{met}, \text{cloud}\} \quad (10)$$

The three error metrics are regarded as equally important, so the overall weight of each path is taken as the arithmetic mean of its three metric-wise weights:

$$w_i = \frac{1}{3} \sum_{m=1}^3 w_i^{(m)} \quad (11)$$

Using the single-path accuracy statistics of the two paths, the metric-wise weights obtained from Eq. (10) are 0.7341 (NMAE), 0.8750 (NMSE) and 0.7262 (NRMSE) in favour of the meteorological-data path. Their arithmetic mean gives $w_{\text{met}} = 0.7784$ and $w_{\text{cloud}} = 0.2216$. Rounding these two estimates to one decimal place—and noting that the two weights must sum to unity—yields the fixed fusion weights $w_{\text{met}} = 0.8$ and $w_{\text{cloud}} = 0.2$. The three metric-wise weights are highly consistent (0.7262–0.8750 for the meteorological-data path), so the averaged result is not driven by any single metric. It should be emphasized that 0.8/0.2 is the one-decimal rounding of the inverse-error estimate (0.7784/0.2216), rather than an arbitrary choice. This outcome is also physically consistent with the nature of the two paths: the meteorological-data path is directly linked to core physical drivers such as irradiance and temperature and is more accurate on every metric, whereas the ground-based cloud image path, although rich in visual dynamic information, suffers from error propagation through its two-stage “predict-then-regress” architecture and is even less accurate than the persistence benchmark; it is therefore the less reliable modality and receives the lower weight.

Table 7 compares the error metrics of the persistence benchmark, the single-modality models, and the proposed decision-level fusion model. As shown in Table 7, the proposed model achieves the lowest NMAE, NMSE, and NRMSE and the highest R^2 among all compared models, confirming that decision-level multimodal fusion significantly outperforms any single-modality forecasting model.

Table 7: Comparison of the error metrics for each model.

Model	NMAE	NMSE	NRMSE	R^2 Score
Persistence forecasting	0.0785	0.0130	0.1141	0.7960
CatBoost based on historical power	0.0691	0.0114	0.1068	0.8186
CatBoost based on meteorological and power data	0.0434	0.0044	0.0662	0.9486
Prediction model based on ground-based cloud images	0.1198	0.0308	0.1756	0.8605
Proposed model	0.0382	0.0037	0.0608	0.9550

4 Conclusion

This chapter focuses on an ultra-short-term photovoltaic power forecasting method based on decision-level multimodal fusion. To overcome the shortcomings of single-source models under complex weather,

we propose a solution that fuses ground-based cloud image visual data with multi-source meteorological physical parameters, and we specifically explore decision-level fusion of the model outputs.

First, two independent single-modality prediction paths are constructed. Path 1 is an indirect prediction model based on ground-based cloud images. This path innovatively adopts a two-stage architecture of “first predicting cloud images, then regressing to power”: in the first stage, a ResSE-STNet model integrating the SE attention mechanism and residual blocks is designed, successfully achieving high-fidelity prediction of future ground-based cloud images (PSNR = 30.34, SSIM = 0.9336); in the second stage, a ResNet network is used to establish an end-to-end mapping from predicted cloud images to power, verifying the feasibility of using visual features for power prediction, with the regression model achieving an R^2 of 0.9652. However, the overall prediction accuracy of the two-stage series connection ($R^2 = 0.8605$) decreases due to the propagation of cloud image prediction errors. Path 2 is a direct prediction model based on CatBoost using meteorological data. This model effectively captures the nonlinear relationships between physical parameters such as irradiance and temperature and power output, achieving an excellent performance with an R^2 of 0.9486. Moreover, feature importance analysis reveals that Power Station Irradiance (PSI) is the most critical predictor.

Second, at the decision level, the prediction outputs of the two paths are combined through weighted fusion with fixed weights. The prediction accuracy of the fusion model significantly surpasses that of any single path, verifying the effectiveness of decision-level fusion.

Despite these promising results, several limitations should be acknowledged. (1) Because PV forecasting is fundamentally an engineering task tightly coupled with plant operation, the exclusion of real-time operational parameters (e.g., inverter status, tracking angles, soiling)—due to data-access constraints—may cap the practically achievable accuracy. (2) The fixed 20%/80% fusion weights, though empirically effective here, still lack a rigorous rationale; systematic comparisons with equal weighting, error-driven adaptive weighting, or validation-set-optimized schemes remain absent. (3) Systematic ablation analyses of key components such as the SE module, residual structure, predicted cloud images, and the fusion rule have so far been confined mainly to the cloud-image generation stage and have not yet been propagated to the downstream power-forecasting task. (4) Validation relies on only a three-month record from a single station, leaving seasonal and cross-site generalizability unverified; (5) the current evaluation is still reported on the whole test set without stratifying samples by representative weather regimes and reporting per-weather-type accuracies for each path and the fusion model separately.

Accordingly, future work will (i) incorporate plant operational variables and physical constraints into the input space; (ii) extend ablation studies from cloud synthesis through to end-to-end forecasting and test adaptive/optimized fusion strategies, thereby quantifying the marginal contribution of each module; and (iii) validate the framework over multi-site, multi-season datasets with weather-type-stratified evaluation to quantify consistent per-regime gains and delineate its robustness and engineering applicability boundaries. These limitations will be systematically addressed in our subsequent research.

Acknowledgement: None.

Funding Statement: This work was financially supported by the Qinghai Provincial Key Research and Development Program (First funding, Grant No. 2025-QY-244): Research on Novel Ultra-short-term Photovoltaic (Solar Thermal) Power Generation Output and Medium- and Long-Term Power Generation Forecasting and Its Application. This work was also supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Second funding, Grant No. 2019QZKK0804): Evaluation of Current Status and Long-term Prospect of Clean Energy.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Yifeng Ma and Wenzheng Yu; methodology, Jianzhou Lin and Yifeng Ma; software, Jianzhou Lin; validation, Yifeng Ma and Zhiyuan You; formal analysis, Jianzhou Lin; investigation, Jianzhou Lin; resources, Yong Xie, Shidong Feng, Zhiyuan You, Zhidan Hui, Zhanyuan Ma, Guoyu Niu and Fa Wen; data curation, Shidong Feng and Yong Xie; writing—original draft preparation, Jianzhou Lin; writing—review and editing, Wenzheng Yu and Yifeng Ma; visualization, Jianzhou Lin; supervision, Wenzheng Yu; project administration, Wenzheng Yu; funding acquisition, Yong Xie, Shidong Feng, Zhidan Hui, Zhanyuan Ma, Guoyu Niu and Fa Wen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Wenzheng Yu, upon reasonable request.

Ethics Approval: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Interpretation
AdamW	Adam with Decoupled Weight Decay
ARMA	Autoregressive Moving Average
CatBoost	Categorical Boosting
DHI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance
GAN	Generative Adversarial Network
GBDT	Gradient Boosting Decision Tree
GHI	Global Horizontal Irradiance
L1	L1 Loss
LSSVM	Least Squares Support Vector Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
NMAE	Normalized Mean Absolute Error
NMSE	Normalized Mean Squared Error
NRMSE	Normalized Root Mean Squared Error
NSRDB	National Solar Radiation Database
PCA	Principal Component Analysis
PSI	Power Station Irradiance
PSNR	Peak Signal-to-Noise Ratio
PSO	Particle Swarm Optimization
PV	Photovoltaic
R^2	Coefficient of Determination
ReLU	Rectified Linear Unit
ResNet	Residual Network
ResSE-STNet	Residual Squeeze-and-Excitation Spatiotemporal Network
RH	Relative Humidity
RMSE	Root Mean Squared Error
SE	Squeeze-and-Excitation
SSIM	Structural Similarity Index
STEP	Second Tibetan Plateau Scientific Expedition and Research Program
SVM	Support Vector Machine
TEMP	Temperature
VAE	Variational Autoencoder
XGBoost	Extreme Gradient Boosting

References

1. Liu R, Li X, Xu P, Li H, Wang L, Tang S. Short-term photovoltaic power forecasting using a reformer-based model with multi-scale dynamic temporal convolution and enhanced invertible instance normalization. *Energy Rep.* 2026;15:109228. [[CrossRef](#)].
2. Li F, Chen S, Ju C, Zhang X, Ma G, Huang W. Research on short-term joint optimization scheduling strategy for hydro-wind-solar hybrid systems considering uncertainty in renewable energy generation. *Energy Strategy Rev.* 2023;50:101242. [[CrossRef](#)].
3. Li J, Ren X, Zhang F, Gao L, Li J, Cui T. Multi-timescale photovoltaic station power prediction based on Reformer model. *Energy.* 2025;327:136423. [[CrossRef](#)].
4. Wang T, Xu Y, Qin Y, Wang X, Zheng F, Li W. Short-term PV forecasting of multiple scenarios based on multi-dimensional clustering and hybrid transformer-BiLSTM with ECPO. *Energy.* 2025;334:137654. [[CrossRef](#)].
5. Wang X, Hu M, Luo X, Guan X. Spatio-temporal photovoltaic power forecasting via Bayesian-optimized dynamic graph convolutional networks with temporal convolutional networks. *J Clean Prod.* 2026;542:147641. [[CrossRef](#)].
6. Das O, Dahlioui D, Hamza Zafar M, Akhtar N, Kumayl Raza Moosavi S, Sanfilippo F. Ultra-short term PV power forecasting under diverse environmental conditions: a case study of Norway. *Energy Convers Manag X.* 2025;27:101072. [[CrossRef](#)].
7. Ma T, Yang H, Lu L. Solar photovoltaic system modeling and performance prediction. *Renew Sustain Energy Rev.* 2014;36:304–15. [[CrossRef](#)].
8. Babatunde AA, Abbasoglu S. Predictive analysis of photovoltaic plants specific yield with the implementation of multiple linear regression tool. *Environ Prog And Sustain Energy.* 2019;38(4):13098. [[CrossRef](#)].
9. Ding M, Xu NZ. A method to forecast short-term output power of photovoltaic generation system based on Markov chain. *Power Syst Technol.* 2011;35(1):152–7. (In Chinese). [[CrossRef](#)].
10. He WG, Hao XJ, Guo YJ, Cao X, Chen JM, Mei F, et al. Ultra-short term power forecast based on ARIMA and SVR for photovoltaic power station. *Guangdong Electr Power.* 2017;30(8):32–7. (in Chinese).
11. Chang GW, Lu HJ. Integrating gray data preprocessor and deep belief network for day-ahead PV power output forecast. *IEEE Trans Sustain Energy.* 2020;11(1):185–94. [[CrossRef](#)].
12. Ding S, Li R, Tao Z. A novel adaptive discrete grey model with time-varying parameters for long-term photovoltaic power generation forecasting. *Energy Convers Manag.* 2021;227:113644. [[CrossRef](#)].
13. Kushwaha V, Pindoriya NM. A SARIMA-RVFL hybrid model assisted by wavelet decomposition for very short-term solar PV power generation forecast. *Renew Energy.* 2019;140:124–39. [[CrossRef](#)].
14. Fara L, Diaconu A, Craciunescu D, Fara S. Forecasting of energy production for photovoltaic systems based on ARIMA and ANN advanced models. *Int J Photoenergy.* 2021;2021:6777488. [[CrossRef](#)].
15. Malvoni M, De Giorgi MG, Congedo PM. Photovoltaic forecast based on hybrid PCA–LSSVM using dimensionality reduced data. *Neurocomputing.* 2016;211:72–83. [[CrossRef](#)].
16. Xiang MJ. Probabilistic forecasting of photovoltaic power generation based on ground-based sky image [dissertation]. Hangzhou, China: Zhejiang University; 2022. (In Chinese). [[CrossRef](#)].
17. Xu Z, Du J, Wang J, Jiang C, Ren Y. Satellite image prediction relying on GAN and LSTM neural networks. In: *Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC)*; 2019 May 20–24; Shanghai, China. [[CrossRef](#)].
18. Acharya SK, Wi YM, Lee J. Day-ahead forecasting for small-scale photovoltaic power based on similar day detection with selective weather variables. *Electronics.* 2020;9(7):1117. [[CrossRef](#)].
19. Zang H, Chen D, Liu J, Cheng L, Sun G, Wei Z. Improving ultra-short-term photovoltaic power forecasting using a novel sky-image-based framework considering spatial-temporal feature interaction. *Energy.* 2024;293:130538. [[CrossRef](#)].
20. Xu S, Zhang R, Ma H, Ekanayake C, Cui Y. On vision transformer for ultra-short-term forecasting of photovoltaic generation using sky images. *Sol Energy.* 2024;267:112203. [[CrossRef](#)].
21. Huang X, Liu J, Xu S, Li C, Li Q, Tai Y. A 3D ConvLSTM-CNN network based on multi-channel color extraction for ultra-short-term solar irradiance forecasting. *Energy.* 2023;272:127140. [[CrossRef](#)].
22. Sun Q, Yan F, Sun W, Zhou Y. DWT-Former: fusing wavelet-based multi-scale features and transformer-based temporal representations for photovoltaic power forecasting. *Energy.* 2025;341:139283. [[CrossRef](#)].

23. Nie Y, Zelikman E, Scott A, Paletta Q, Brandt A. SkyGPT: probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained VideoGPT. *Adv Appl Energy*. 2024;14:100172. [[CrossRef](#)].
24. Ruan G, Chen X, Lim EG, Fang L, Su Q, Jiang L, et al. On the use of sky images for intra-hour solar forecasting benchmarking: comparison of indirect and direct approaches. *Sol Energy*. 2024;276:112649. [[CrossRef](#)].
25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018 Jun 18–23; Salt Lake City, UT, USA. [[CrossRef](#)].
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. [[CrossRef](#)].
27. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*; 2015 May 7–9; San Diego, CA, USA.
28. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*; 2019 May 6–9; New Orleans, LA, USA.
29. Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*; 2017 Apr 24–26; Toulon, France.
30. Micikevicius P, Narang S, Alben J, Diamos G, Elsen EK, Garcia D, et al. Mixed precision training. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*; 2018 Apr 30–May 3; Vancouver, BC, Canada.
31. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*; 2013 Jun 16–21; Atlanta, GA, USA.
32. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Boca Raton, FL, USA: CRC Press; 1984. [[CrossRef](#)].
33. Quinlan JR. *C4.5: Programs for machine learning*. San Mateo, CA, USA: Morgan Kaufmann; 1993.
34. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. [[CrossRef](#)].
35. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*; 2018 Dec 3–8; Montréal, QC, Canada.
36. Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The national solar radiation data base (NSRDB). *Renew Sustain Energy Rev*. 2018;89:51–60. [[CrossRef](#)].
37. Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM. In: *Proceedings of the 2010 20th International Conference on Pattern Recognition*; 2010 Aug 23–26; Istanbul, Turkey. [[CrossRef](#)].
38. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600–12. [[CrossRef](#)].
39. Ma Y, Yu W, Zhu J, You Z, Jia A. Research on ultra-short-term photovoltaic power forecasting using multimodal data and ensemble learning. *Energy*. 2025;330:136831. [[CrossRef](#)].
40. Xiong B, Chen Y, Zhao X, Su Z, Fu J, Chen D, et al. Multimodal ultra-short-term probabilistic solar power forecasting with generative AI and Transformer. *Adv Appl Energy*. 2025;20:100250. [[CrossRef](#)].