SolarFusionNet: Enhanced Solar Irradiance Forecasting via Automated Multi-Modal Feature Selection and Cross-Modal Fusion

Tao Jing, *Student Member, IEEE*, Shanlin Chen, David Navarro-Alarcon, *Senior Member, IEEE*, Yinghao Chu and Mengying Li

***Abstract*—Solar forecasting has emerged as a cost-effective technology to mitigate the negative impacts of intermittent solar power on the power grid. Despite the multitude of deep learning methodologies available for forecasting solar irradiance, there is a notable gap in research concerning the automated selection and holistic utilization of multi-modal features for ultra-short-term regional irradiance forecasting. Our study introduces SolarFusionNet, a novel deep learning architecture that effectively integrates automatic multi-modal feature selection and cross-modal data fusion. SolarFusionNet utilizes two distinct types of automatic variable feature selection units to extract relevant features from multichannel satellite images and multivariate meteorological data, respectively. Long-term dependencies are then captured using three types of recurrent layers, each tailored to the corresponding data modal. In particular, a novel Gaussian kernel-injected convolutional long short-term memory network is specifically designed to isolate the sparse features present in the cloud motion field derived from optical flow. Subsequently, a hierarchical multi-head cross-modal self-attention mechanism is proposed based on the physical-logical dependencies among the three modalities to investigate the coupling correlations among the modalities. The experimental results indicate that SolarFusionNet exhibits robust performance in predicting regional solar irradiance, achieving higher accuracy than other state-of-the-art models and a forecast skill ranging from 37.4% to 47.6% against the smart persistence model for the 4-hour-ahead forecast.**

***Index Terms*—Solar irradiance forecasting, Multi-modal deep learning, Attention mechanism, Optical flow**

I. Introduction

THE deployment of solar technologies, especially photovoltaic (PV), has increased significantly in recent years due to concerns about global climate change, supportive government policies, and lower equipment costs. Despite the promise of solar energy, the main challenge to its feasibility is its highly volatile and intermittent nature. Consequently, accurate prediction of solar irradiance and power production has become a crucial requirement for stable operation of the electrical grid [1], [2]. Current solar forecasting applications focus mainly on solar irradiance forecasting and solar power forecasting [3]. Solar power forecasting is calculated from irradiance prediction data and a range of possible predictors based on regressive models or model chains [4], [5]. Therefore, solar irradiance prediction is fundamental for solar power forecasting. In this research, we focus on global horizontal irradiance (GHI) prediction. The GHI is predominantly modulated by clouds, aerosols, and water vapor through a sophisticated process of radiative transfer in the atmosphere. The spatial and temporal variability of these factors, especially cloud fields, makes the prediction of GHI an exceptionally demanding endeavor [6]. Consequently, accurately capturing cloud motions is essential for reliable GHI forecasting.

In recent years, deep learning techniques have gained widespread attention among solar engineers due to robust generalizability, efficient handling of unstructured data, and automated feature extraction [7], [8]. With the availability of a wide range of satellite data, all-sky images, numerical weather predictions (NWP), historical meteorological data, etc., deep learning models for cloud dynamics extraction and solar forecasting show promising performance [9]. Although all-sky images provide information on small-scale cloud cover dynamics, satellite images provide not only information on local cloud cover dynamics but also on the spatial dynamics of neighboring regions, providing a robust data infrastructure for solar irradiance forecasting based on deep learning models [6], [10], [11].

However, the efficient utilization of multi-modal data sources to construct accurate prediction models is facing significant challenges. Multi-modal data provide a huge number of input features, which have complicated nonlinear relationships with the prediction targets. Although the inclusion of certain features can be beneficial in improving the accuracy of the prediction, the input of redundant information not only wastes computational resources but also negatively affects the accuracy of the prediction [12]. Existing research generally tends to apply statistical methods for feature selection, and then input the filtered features to deep learning models [7], [12], [13], [14]. For example, Nejati et al. [12] calculated the correlation factor between the input meteorological variables and solar irradiance based on the theory of mutual information (MI) to predict solar power. Bouzgou et al. [13] proposed a Wrapper Mutual Information Methodology (WMIM) that integrates Extreme Learning Machine (ELM) and MI to

This work is substantially supported by three grants from the Research Grants Council of the Hong Kong (Project No. C6003-22Y, No. 21200424, and No. 25213022) and a grant from the Guangdong Basic and Applied Basic Research Foundation (Project No. 2024A1515010117). *Corresponding authors: Mengying Li and Yinghao Chu.*

T. Jing, D. Navarro-Alarcon and M. Li are with Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong. Email: tao-joty.jing@connect.polyu.hk, dnavar@polyu.edu.hk, mengying.li@polyu.edu.hk

S. Chen and Y. Chu are with Department of Systems Engineering, City University of Hong Kong (CityU), Kowloon, Hong Kong. Email: shanchen@cityu.edu.hk, yinghchu@cityu.edu.hk

predict solar irradiance. This methodology performs feature selection by optimizing the similarity function between input variables. However, these statistically based feature selection methods have some obvious limitations when used as pre-processing steps in deep learning. Firstly, statistical-based analysis methods usually rely on global features for similarity calculations, which ignores local similarities between covariate features and target features, leading to information redundancy or inefficient use. Secondly, when the results of feature selection are used as input for deep learning models, statistically based feature selection methods do not adequately capture the complex non-linear relationships between covariate features and predictive targets. This oversight can lead to excessive redundancy or insufficient information in the feature selection process, ultimately compromising prediction accuracy.

Moreover, the effective fusion of multi-modal features is also one of the significant challenges for solar irradiance forecasting [15]. Complex coupling correlations are presented among the multi-modal features. For example, optical flow signals derived from the visible and infrared channels of satellite images reveal cloud trajectories and dynamic density information, respectively, which are essential for refining cloud dynamics based on satellite images [16]. Therefore, it is essential to extract optical flow features as a distinct mode and integrate them with multi-channel satellite images and multi-variable meteorological data. However, there are fewer researches utilizing optical flow features derived from satellite images as input sources, and there are research gaps regarding algorithms for sparsity feature extraction of optical flow signals at high temporal resolution. Although Boussif et al. [11] and Liu et al. [17] utilized optical flow signals derived from satellite images and all-sky images to predict solar irradiance, the research did not explore in-depth the critical role of optical flow signals in extracting cloud motion features based on satellite images.

Two typical multi-modal feature fusion techniques are included in deep learning-based solar forecasting: linear aggregation [18], [19], and cross-attention mechanism fusion [11], [17]. Ajith et al. [19] used a fully connected layer to concatenate extracted features from infrared sky images as well as historical GHI. However, exploring the coupling correlation between multi-modal features with linear aggregation is challenging. Cross-attention mechanisms have excellent global information search capabilities, which provide excellent techniques for mining coupled correlations. Liu et al. [17] proposed a multi-modal information fusion-based framework to encode historical clear-sky GHI and all-sky images, subsequently, a cross-modal attention mechanism was used to explore the coupled correlations between the two modalities. Boussif et al. [11] used the Crossformer architecture to combine satellite data features and ground-based measurements for the prediction of day-ahead solar radiation. However, vanilla cross-attention techniques are typically used to fuse two modal features. When optical flow signals are involved in the fusion process as a distinct mode, the coupling correlation becomes more complicated. Theoretically, the input data of the three modalities used in this work revealed the following physical logical dependencies: first, the sparse optical flow signals enable the predictive model to accurately identify the cloud motion features in satellite images over time; and second, the infrared and visualization channels of the satellite images jointly provide the shape and structure features of the clouds, which complement the deficiencies of ground-based sensors in collecting cloud information. Therefore, how to effectively integrate optical flow signals, satellite image data, and meteorological data, and accurately model the physical logical dependencies between multi-modal features, posing a challenge for multi-modal feature fusion.

To bridge the aforementioned research gaps, we propose an end-to-end deep learning model called SolarFusionNet based on the self-attention mechanism for automatic selection of multi-modal features and cross-modal features fusion to enhance regional ultra-short-term (i.e. from 10 minutes to 4 hours ahead) solar irradiance prediction. To demonstrate the validity of SolarFusionNet, we performed experiments at four Baseline Surface Radiation Network (BSRN) stations [20], and compared the prediction results with the state-of-the-art (SOTA) benchmark models. The main contributions are:

- We propose a novel automatic multi-modal feature selection and fusion framework to deeply explore the coupled correlations among different modal features. The model incorporates two automatic multi-modal feature selection mechanisms, three specific Recurrent Neural Networks (RNNs) for spatio-temporal feature encoding, and an attention mechanism-based cross-modal fusion strategy, aiming to enhance the accuracy of regional solar irradiance prediction with high time resolution (i.e. 10-minute resolution).
- To minimize information redundancy and maximize the impact of relevant input variables, we introduce specialized Meteorological Selection Units (MSU) and Spectral Selection Units (SSU). Utilizing the local information extraction capability of convolutional operation and the gating mechanism, the units are able to effectively maximize the weights of the relevant input variables and minimize the weights of redundant information.
- Optical flow signals derived from high time resolution satellite images exhibit sparsity. To efficiently extract features from sparse optical flow signals, we propose a Gaussian Kernel-injected Convolutional Long Short-Term Memory Network (GKConvLSTM). Adaptive kernel weights are then computed using the normalized local density values and a predefined Gaussian Kernel (GK).
- A hierarchical multi-head self-attention mechanism is proposed based on the physical-logical dependencies among the three modalities for cross-modal feature cross-fertilization and eliminates redundant information. Such an approach aims to utilize the complementary strengths of various features to enhance the overall prediction performance.

The remainder of the paper is organized as follows. Section II describes the multi-horizon GHI forecasting procedure and the multimodal data pre-processing method. Section III presents the proposed SolarFusionNet. Experimental details and performance evaluation are discussed in Section IV. Finally, Section V provides the conclusion.

II. Preliminaries

The focal objective of this work is the development of an integrated framework for solar irradiance forecasting, utilizing optical flow derived from satellite images, original multi-channel satellite images and historical multivariate meteorological data from various locations. In this section, we elucidate the mathematical issues of multi-horizon GHI prediction utilizing multi-modal data, as well as the pre-processing techniques employed in the study.

A. Multi-Horizon GHI Forecasting with Multi-modal Data

In GHI forecasting scenarios, the GHI clear sky index (CSI) is usually selected as the primary prediction target to remove the influences of seasonal patterns, thus improving the accuracy of the forecast in a range of predictive methodologies [5]. The CSI is calculated as:

$$\mathrm{CSI}^t = \frac{\mathrm{GHI}^t}{\mathrm{GHI}_{\mathrm{cs}}^t}, \tag{1}$$

where $\mathrm{GHI}_{\mathrm{cs}}$ refers to the GHI value under clear sky condition, which can be derived from a physical or empirical clear-sky model. Here, we use the Perez-Ineichen model to calculate $\mathrm{GHI}_{\mathrm{cs}}$ [21], and set the CSI to 1 at night when $\mathrm{GHI}_{\mathrm{cs}} = 0$.

Satellite images, optical flow signals derived from satellite images, and ground measurements of various meteorological data are utilized as inputs. Satellite images and optical flow signals can provide information on cloud movements, water vapor, and aerosol levels that are often inaccessible through ground measurement stations. Thus, it is imperative to develop a multi-modal data fusion model $f_q(\cdot)$ to effectively fuse satellite images $\boldsymbol{\mathcal{X}}_S \in \mathbb{R}^{T \times C_s \times H \times W}$, optical flow signals $\boldsymbol{\mathcal{X}}_{Of} \in \mathbb{R}^{(T-1) \times C_{Of} \times H \times W}$ and meteorological data $\boldsymbol{\mathcal{X}}_M \in \mathbb{R}^{T \times C_{ts}}$ to enhance the accuracy of end-to-end GHI predictions. Each batch forecast takes the form:

$$\hat{\mathcal{Y}}_i(t,\tau) = f_q(\tau, \mathcal{Y}_i^{t-k:t}, \boldsymbol{\mathcal{X}}_{S_i}^{t-k:t}, \boldsymbol{\mathcal{X}}_{Of_i}^{t-k:t}, \boldsymbol{\mathcal{X}}_{M_i}^{t-k:t}, \boldsymbol{\mathcal{S}}_i), \tag{2}$$

where, $\hat{\mathcal{Y}}_i(t,\tau)$ indicates the predicted CSI of the τ-step-ahead forecast at time t, $\mathcal{Y}_i^{t-k:t}$ is the k-step labels of input batch, $\boldsymbol{\mathcal{S}}_i$ indicates spatial information including the longitude, latitude, and altitude of each location, as well as the geospatial data within the satellite image coverage area. In line with other direct methods, we simultaneously output forecasts for τ_{max} time steps (i.e., $\tau \in \{1, \cdots, \tau_{\max}\}$). We integrate all multi-modal historical information within a finite look-back window k, using CSI and known inputs only up to and including the forecast start time t (i.e., $\mathcal{Y}_i^{t-k:t} = \{\mathcal{Y}_i^{t-k}, \cdots, \mathcal{Y}_i^{t}\}$). The predicted CSI values are then converted back to GHI for evaluation using Eq.(1) with $\mathrm{GHI}_{\mathrm{cs}}^{t+\Delta t}$.

B. Data Description and Pre-processing

The following two types of datasets are utilized:

1) Time Series Data: Time-series of measured GHI, Beam Normal Irradiance (BNI), Diffuse Horizontal Irradiance (DHI), and meteorological parameters such as temperature, relative humidity, and atmospheric pressure, were collected at 10-minute resolution over 7-year (2016-2022) from four BSRN stations [20], as shown in Fig. 1.

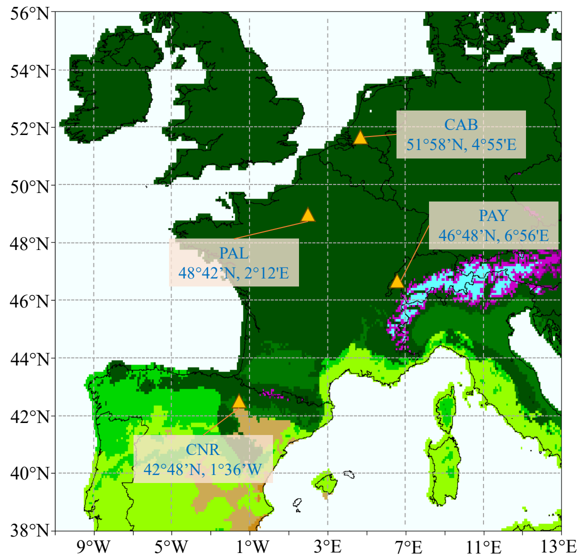


Fig. 1. Geographical distribution of the four selected BSRN stations (triangle symbols) in the updated Köppen–Geiger climate classification system [22].

To ensure data quality and the robustness and reliability of the deep learning model, rigorous quality control (QC) on the collected measurements is performed using two filters (Extremely-rare limits & Closure equation) [23]:

$$\begin{bmatrix} -2 < \mathrm{GHI} < 1.2 E_{0n} \cos^{1.2}(Z) + 50 \\ -2 < \mathrm{DHI} < 0.75 E_{0n} \cos^{1.2}(Z) + 30 \\ -2 < \mathrm{BNI} < 0.95 E_{0n} \cos^{0.2}(Z) + 10 \end{bmatrix}$$
$$\begin{bmatrix} |\mathrm{closr}| < 8\% & \text{for } Z < 75^\circ \text{ and GHI} > 50 \\ |\mathrm{closr}| < 15\% & \text{for } 75^\circ < Z < 93^\circ \text{ and GHI} > 50 \end{bmatrix}.$$

In the above QC procedure, Z is solar zenith angle, E_{0n} is extraterrestrial irradiance on a surface normal to the solar ray, $|\mathrm{closr}| = |\mathrm{GHI} - (\mathrm{DNI} \cos Z + \mathrm{DHI})|$ is the difference between measured and computed GHI.

2) Satellite Images: Multi-channel satellite images are obtained from the Meteosat Second Generation Rapid Scan Service (MSG-RSS) operated by EUMETSAT [24]. We selected Rectified RSS images (level 1.5), which have spatial coverage spanning a longitudinal range from -65° to 84°, and a latitudinal range from 16° to 70°. The data product consists of satellite images collected in 12 spectral wavelength channels (8 in the thermal infrared spectrum, 3 in the visible spectrum, and 1 in the near-infrared spectrum, as shown in Table I). HRV represents the High Resolution Visible Channel, distinguished by its precise spatial resolution of 1 km. The remaining 11 channels are low-resolution channels with the spatial resolution of 3 km. Therefore, the HRV channel is excluded in this study to maintain the consistency of spatial resolution across different satellite image channels.

Satellite images with 10-minute resolution of VIS 0.6, VIS 0.8, WV 6.2, WV 7.3, and IR 10.8 are selected as the inputs. The two channels in the visible spectrum, VIS 0.6 and VIS 0.8, could provide cloud images during daytime. The chosen wavelengths allow the distinction from the Earth's surface of different cloud types, as well as support the determination of the atmospheric aerosol content. The two channels in the water-vapour absorption band, WV 6.2 and WV 7.3, provide the water-vapour distribution for two distinct layers

TABLE I
AN OVERVIEW OF THE 12 SEVIRI CHANNELS [24]

Channel	Absorption Band	Wavelength (μm)	Bandwidth (μm)
HRV	Visible	0.75	0.6-0.9
VIS 0.6	VNIR	0.635	0.56-0.71
VIS 0.8	VNIR	0.81	0.74-0.88
IR 1.6	VNIR	1.64	1.50-1.78
IR 3.9	Window	3.92	3.48-4.36
WV 6.2	Water Vapor	6.25	5.35-7.15
WV 7.3	Water Vapor	7.35	6.85-7.85
IR 8.7	Window	8.70	8.30-9.10
IR 9.7	Ozone	9.66	9.38-9.94
IR 10.8	Window	10.80	9.80-11.80
IR 12.0	Window	12.00	11.00-13.00
IR 13.4	Carbon Dioxide	13.40	12.40-14.40

in the troposphere. These two channels can also be used to derive atmospheric motion vectors in cloud-free areas, and will support the IR 10.8 channel in the determination of the height of semitransparent clouds [25], [26]. We initially trim the satellite images to encompass a latitudinal span from -6.64° to 10.51° and a longitudinal stretch from 38.82° to 55.97° to cover the four BSRN sites, and reshape satellite image size to 64×64. To facilitate our analysis, we convert the satellite imagery from a geostationary projection to the World Geodetic System 1984 (WGS 84) coordinate frame [11] and perform standard deviation normalization in every channel.

III. METHODOLOGY

To address the challenges of ramp events in the GHI prediction process, SolarFusionNet employs an automated feature selection strategy, augmented by an attention-driven mechanism, to adeptly capture the cloud motion from satellite images. The integration of these elements is crucial to increase the accuracy of GHI predictions. Figure 2 illustrates the framework of SolarFusionNet. First, we develop two distinct feature selection networks: SSU and MSU, which are utilized to assess and prioritize the importance of input features from multi-spectral satellite images $\boldsymbol{\mathcal{X}}_S \in \mathbb{R}^{T\times C_s\times H\times W}$ and meteorological data $\boldsymbol{\mathcal{X}}_M \in \mathbb{R}^{T\times C_{ts}}$, respectively. Furthermore, the $\mathcal{TV}-\mathcal{L}1$ algorithm [27] is utilized to derive optical flow signals for each spectral channel, increasing spatial background details. Each optical flow channel shares the same weight $\mathcal{W}_t$ as the corresponding satellite image channel. Each dataset is encoded using encoders customized for its respective modality. Meteorological features are encoded using the vanilla long-short-term memory (LSTM) network [28], adept at capturing temporal dynamics. Satellite images benefit from the spatial-temporal capabilities of the vanilla Convolutional Long-Short-Term Memory (ConvLSTM) network [29], which excels at interpreting visual patterns over time. Optical flow signals are accurately encoded by a specialized GKConvLSTM, which is specifically designed to integrate the nuances of motion and spatial features. To effectively fuse multimodal features, we propose a hierarchical multi-head cross-modal self-attention mechanism. The following subsections describe in detail the customized modules of SolarFusionNet.

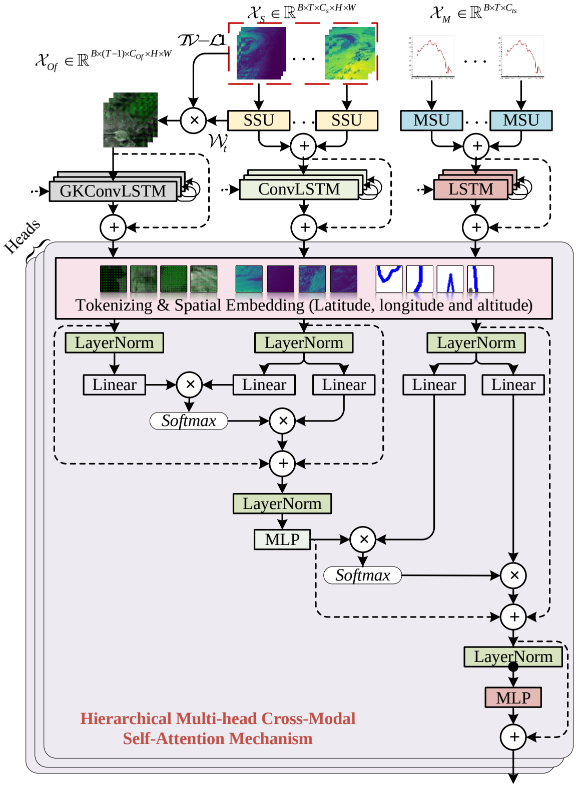

$\mathcal{X}_S \in \mathbb{R}^{B\times T\times C_s\times H\times W}$
$\mathcal{X}_M \in \mathbb{R}^{B\times T\times C_{ts}}$
$\mathcal{X}_{Of} \in \mathbb{R}^{B\times(T-1)\times C_{Of}\times H\times W}$
$\mathcal{TV}-\mathcal{L}1$
$\mathcal{W}_t$
SSU
SSU
MSU
MSU
GKConvLSTM
ConvLSTM
LSTM
Heads
Tokenizing & Spatial Embedding (Latitude, longitude and altitude)
LayerNorm
LayerNorm
LayerNorm
Linear
Linear
Linear
Linear
Linear
Softmax
LayerNorm
MLP
Softmax
LayerNorm
MLP
Hierarchical Multi-head Cross-Modal Self-Attention Mechanism

Fig. 2. SolarFusionNet: Attention-driven multi-modal feature fusion framework for GHI forecasting.

A. Automated Multi-modal Feature Selection Units

The intricate correlations between various meteorological data and satellite channels are typically elusive, making it challenging to anticipate which variables are significant. To empower the model with the ability to dynamically and autonomously process multivariate features in a non-linear manner, we propose two automatic feature selection units, MSU and SSU. Inspired by the Temporal Fusion Transformer (TFT) [30], the MSU and SSU, as depicted in Fig. 3, are designed to process historical meteorological data and satellite spectral channels, respectively. The units aim to improve the performance of the model by enabling it to identify and utilize the most informative features without manual intervention. In addition to clarifying which variables are significant for the prediction task, the feature selection process enables the model to eliminate extraneous inputs that can introduce noise and adversely impact the prediction performance.

Without loss of generality, we present MSU as follows (Fig. 3 (a)). Let $\boldsymbol{\mathcal{S}}_w^i \in \mathbb{R}^{1\times w}$ denote the transformed input of the ith meteorological feature and w denote the length of the input window, with $\sum_1^{\ell_{\max}^m} \boldsymbol{\mathcal{S}}_w^i$ being the flattened vector of the meteorological features input $\ell_{\max}$. Each input vector $\boldsymbol{\mathcal{S}}_w^i$ is fed through the designed 1-Dimensional Convolutional Residual Unit (CRU$_{1\text{D}}$) which performs the extraction of local temporal features from meteorological data, effectively avoiding the omission of information.

$$\text{CRU}_{1\text{D}}(\boldsymbol{\mathcal{S}}_w^i) = \text{LN}(\sigma(\text{conv}_{1\text{D}}(\varepsilon_1)) \odot \text{conv}_{1\text{D}}(\varepsilon_1) + \boldsymbol{\mathcal{S}}_w^i) \quad (3)$$

$$\varepsilon_1 = \text{conv}_{1\text{D}}(\text{ELU}(\text{conv}_{1\text{D}}(\boldsymbol{\mathcal{S}}_w^i))), \quad (4)$$

where, ELU is the Exponential Linear Unit (ELU) activation function [31], $\varepsilon_1 \in \mathbb{R}^w$ indicates output of intermediate layer, LN represents standard layer normalization [32], $\sigma(\cdot)$ is the $sigmoid$ activation function, $\odot$ is the element-wise Hadamard product. To enhance modeling flexibility, we implemented gated residual connections to selectively suppress any unnecessary components. By employing a $sigmoid$ activation function, it is effective to suppress input features with minimal or no contribution, while excluding extraneous inputs that potentially introduce noise and negatively affect the prediction performance. During training, dropout is applied before the gating layer and LN.

Each $\boldsymbol{\mathcal{S}}_w^i$ is fed into the CRU$_{\text{1D}}$ for feature encoding in the time window w of each input variable. Concurrently, an aggregate of $\sum_1^{\ell_{\max}} \boldsymbol{\mathcal{S}}_w^i$ across all levels up to CRU$_{\text{1D}}$ undergoes a similar encoding process. Subsequently, a $softmax$ layer is employed to assign trainable weights $\boldsymbol{\mathcal{V}}_w$ to each meteorological feature, and the feature filtering is completed by element-wise Hadamard product, which helps establish a nonlinear relationship with the target feature (CSI). The mathematical representation is given by the following expression:

$$\boldsymbol{\mathcal{V}}_w = \text{Softmax}(\text{CRU}_{\text{1D}}(\sum\nolimits_1^{\ell^m_{\max}} \boldsymbol{\mathcal{S}}_w^i)) \tag{5}$$

$$\tilde{\boldsymbol{\mathcal{S}}}_w^i = \text{CRU}_{\text{1D}}^{\tilde{\boldsymbol{\mathcal{S}}}_i}(\boldsymbol{\mathcal{S}}_w^i) \tag{6}$$

$$\tilde{\boldsymbol{\mathcal{S}}}_w = \sum_{i=1}^{\ell^m_{\max}} \mathcal{V}_w^i \tilde{\boldsymbol{\mathcal{S}}}_w^i. \tag{7}$$

For SSU, the fundamental architecture of the MSU is retained; however, to capture the temporal feature and spatial feature nuances of each spectral channel within the input window, we replace CRU$_{\text{1D}}$ in the MSU with CRU$_{\text{3D}}$. In addition, a global average pooling (GAP) layer is utilized to distill spatial information more effectively before performing the $softmax$ operation. The framework of the SSU is depicted in Fig. 3 (b), the mathematical representation of the output, derived from the features processed by the SSU, is given by the following expression:

$$\text{CRU}_{\text{3D}}(\boldsymbol{\mathcal{B}}_w^j) = \text{LN}(\sigma(\text{conv}_{\text{3D}}(\boldsymbol{\eta}_1)) \odot \text{conv}_{\text{3D}}(\boldsymbol{\eta}_1) + \boldsymbol{\mathcal{B}}_w^j) \tag{8}$$

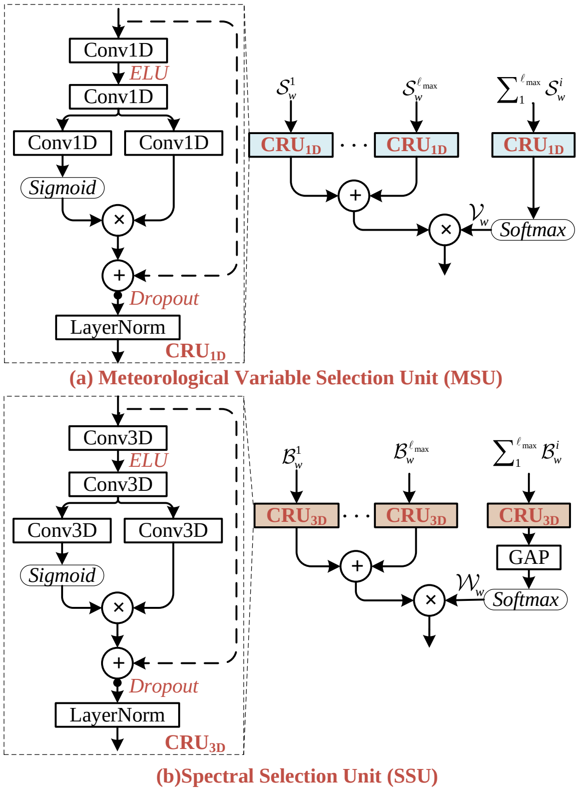
$$\boldsymbol{\eta}_1 = \text{conv}_{\text{3D}}(\text{ELU}(\text{conv}_{\text{3D}}(\boldsymbol{\mathcal{B}}_w^j))) \tag{9}$$

$$\boldsymbol{\mathcal{W}}_w = \text{Softmax}(\text{GAP}(\text{conv}_{\text{3D}}(\sum\nolimits_1^{\ell^b_{\max}} \boldsymbol{\mathcal{B}}_w^j))) \tag{10}$$

$$\tilde{\boldsymbol{\mathcal{B}}}_w^j = \text{CRU}_{\text{3D}}^{\tilde{\boldsymbol{\mathcal{B}}}_j}(\boldsymbol{\mathcal{B}}_w^j) \tag{11}$$

$$\tilde{\boldsymbol{\mathcal{B}}}_w = \sum_{j=1}^{\ell^b_{\max}} \mathcal{W}_w^j \tilde{\boldsymbol{\mathcal{B}}}_w^j. \tag{12}$$

The meaning of each symbol is the same as for MSU.



(a) Meteorological Variable Selection Unit (MSU)

(b)Spectral Selection Unit (SSU)

Fig. 3. (a) Meteorological Variable Selection Unit; (b) Spectral Selection Unit.

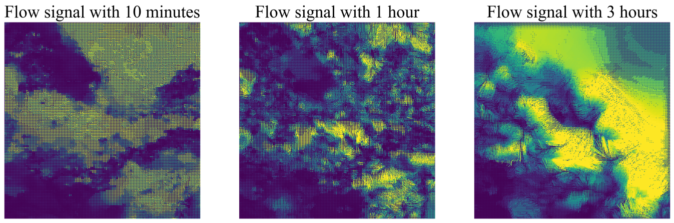


Fig. 4. Optical flow signals with varying time intervals. In cases of high time resolution, the cloud moves over a shorter distance or may not move at all, resulting in sparse optical flow signals. As the time intervals increase, the sparsity of the optical flow information gradually reduces.

B. Gaussian Kernal Injection Convolutional Long Short-Term Memory Network

We observe that the optical flow signals derived from high-temporal-resolution satellite images demonstrates sparsity. As shown in Fig. 4, the sparsity becomes more pronounced as the time intervals shorten. In the feature extraction process from optical flow signals using the vanilla ConvLSTM, the conventional convolution operation struggle to accurately extract key features due to the limitation of local receptive fields when extracting sparse features. Given that high time resolution (10 minutes) cloud motion information primarily exhibits characteristics similar to a low-frequency signal, which can lead to suboptimal performance. To address the challenge, we introduce a novel GKConvLSTM, as shown in Fig. 5.

The GK is well-known as a low-pass filter with anti-aliasing properties, proficient at smoothing out high-frequency information. Consequently, we apply a postprocessing step using a predefined GK after the convolution operation within the ConvLSTM framework which can attenuate high-frequency noise and thus smooth the input. To address the challenge of sparsity and enhance computational efficiency, we develop an adaptive GK weighting algorithm that dynamically adjusts based on the feature map extracted from the pre-trained model. The approach solves the sparsity problem by adjusting the weight of the kernel according to the local density of the feature map.

The size and standard deviation σ of GK are deterministic functions. A two-dimensional GK is formed by sampling a Gaussian distribution in both dimensions:

$$k(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{13}$$

where $k(x, y)$ indicates the x and y spatial dimensions in the kernels. While the direct application of a predefined GK to convolutional outputs is generally effective, which exhibits inefficiencies when applied to sparse optical flow signals [33]. To overcome this limitation, we leverage feature maps obtained from a pre-trained model to calculate local densities. Subsequently, the adaptive kernel weights are calculated using the normalized local density values with the predefined GK. Such an adaptive filtering approach relies on the local density values, thereby ensuring more efficient processing of sparse optical flow information. The mathematical expressions are:

$$D(p_{i,j}) = \frac{1}{k_s^2} \sum_{u=0}^{k_s-1} \sum_{v=0}^{k_s-1} f\left(i+u, j+v\right) \tag{14}$$

$$\gamma_i = \frac{D\left(p_{i,j}\right) - D_{\min}}{D_{\max} - D_{\min}} \tag{15}$$

$$\theta_{G_\sigma} = \gamma_i \cdot k(x, y), \tag{16}$$

where $D(p_{i,j})$ represents the local density values of the feature map $f\left(i+u, j+v\right)$, γ_i is the standardized GK weights, θ_{G_σ} is the adaptive GK with standard deviation σ. The pre-trained model used in this study is ResNet50 [34]. The mathematical formula for the convolution operation in GKConvLSTM is:

$$h_i = \mathrm{ELU}(\mathrm{pool}(\boldsymbol{\theta}_{G_\sigma} \circledast (\boldsymbol{\theta}_w \circledast \boldsymbol{x}_i))), \tag{17}$$

where $\circledast$ represents the convolution operation, $\boldsymbol{\theta}_w$ is the weight of convolutional layer, $\boldsymbol{x}_i$ represents the input tensor.

C. Hierarchical Multi-head Cross-Modal Self-Attention Mechanism

Optical flow and spectral satellite data, along with meteorological variables, have been processed by temporal feature extraction using GKConvLSTM, ConvLSTM, and LSTM, respectively. However, effective integration of features from various data sources presents a pressing challenge. To effectively integrate features, we propose a hierarchical multi-head cross-modal self-attention mechanism. Satellite, optical flow, and meteorological features require patching and embedding of spatial location information before feature fusion. Rotary Positional Embedding (RoPE) [35] is used to embed spatial location information, encompassing latitude, longitude, and altitude for each pixel of satellite images, as well as the location information of each ground-level station.

The process of cross-modal feature fusion is conducted in two primary stages. Initially, for the physical-logical relationship between the optical flow features and the satellite image features, the linear projection of the optical flow features is used as a query ($\boldsymbol{Q}_{of}$), where each $\boldsymbol{Q}_{of}$ maps the dynamic properties in a specific time and region. Meanwhile, the keys ($\boldsymbol{K}_s$) and values ($\boldsymbol{V}_s$) are the linear projections of the satellite image features, which provide spatial cloud information. Each $\boldsymbol{Q}_{of}$ is compared with all $\boldsymbol{K}_s$ under the self-attention mechanism to compute the attention scores between the features of the optical flow and the features of the satellite image. By normalizing the attention scores using $softmax$, these are subsequently used to weight the corresponding $\boldsymbol{V}_s$. This means that the final output $\widetilde{\boldsymbol{\mathcal{H}}}_1$ not only contains information about the optical flow, but also incorporates highly correlated features of the satellite image. In this manner, SolarFusionNet is able to model the cloud information reflected between the optical flow features and the satellite image features. Equations (18) and (19) present the mathematical formulation for the first stage.

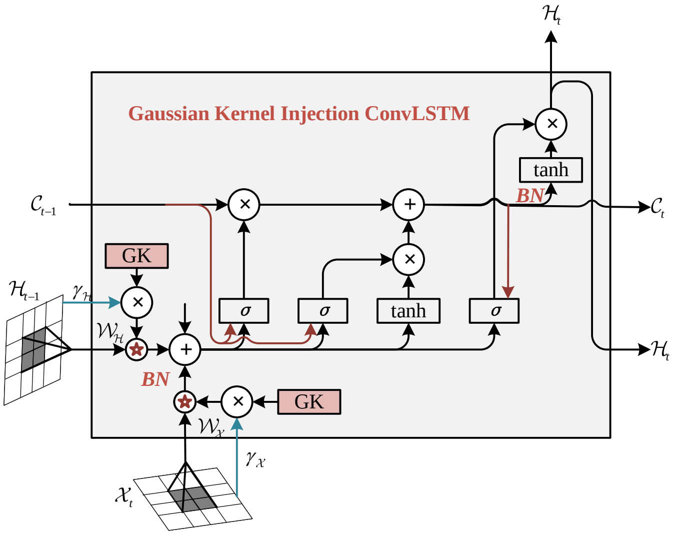


Fig. 5. Gaussian Kernel Injection ConvLSTM.

Subsequently, the integrated output from the first stage is further fused with the features derived from meteorological variables to model the second physical-logical relationship. Specifically, a linear projection of $\widetilde{\boldsymbol{\mathcal{H}}}_1$ is used as the query vector, while the keys ($\boldsymbol{K}_m$) and the values ($\boldsymbol{V}_m$) are generated from meteorological features. The purpose of $\widetilde{\boldsymbol{\mathcal{H}}}_1$ is to compare with each $\boldsymbol{K}_m$ as a mechanism to assess the consistency of the spatial dynamics of each particular region observed from satellites with ground-based meteorological data in the related region. By calculating the dot product between $\widetilde{\boldsymbol{\mathcal{H}}}_1$ and $\boldsymbol{K}_m$, attention scores that reflect the correlation can be obtained. The score is normalized by $softmax$ to obtain the weights corresponding to each $\boldsymbol{K}_m$. Using the weights, the corresponding $\boldsymbol{V}_m$ (i.e., meteorological features) are weighted and summed to produce a composite output. With the application of the self-attention mechanism, SolarFusionNet is able to automatically identify and emphasize the features that are most critical to predicting solar irradiance, thus improving the accuracy and efficiency of prediction. Equation (20) presents the mathematical formulation of the algorithm for the second stage.

Each phase of the feature fusion process utilizes a modified version of the self-attention mechanism [30] to enhance the integration of information. A unified approach [30] is utilized in which each head calculates the attention using identical information, with the outputs subsequently aggregated additively. For the first phase:

$$\text{Atten}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{attn}}}\right)\boldsymbol{V} \tag{18}$$

$$\widetilde{\boldsymbol{\mathcal{H}}}_1 = \frac{1}{m_h}\sum_{h=1}^{m_h}\text{Atten}(\boldsymbol{Q}_{of}\boldsymbol{W}_{Q_{of}}^{(h)}, \boldsymbol{K}_s\boldsymbol{W}_{K_s}^{(h)}, \boldsymbol{V}_s\boldsymbol{W}_{V_s}) \cdot \boldsymbol{W}_{\widetilde{\mathcal{H}}_1}, \tag{19}$$

where $\boldsymbol{W}_{Q_{of}}^{(h)}$ and $\boldsymbol{W}_{K_s}^{(h)}$ are head-specific weights for keys and queries, $\boldsymbol{W}_{V_s}$ are value weights shared across all heads, $\boldsymbol{W}_{\widetilde{\mathcal{H}}_1}$ are used for final linear mapping of first phase. Similarly, for the second stage:

$$\widetilde{\boldsymbol{\mathcal{H}}}_2 = \frac{1}{m_h}\sum_{h=1}^{m_h}\text{Atten}(\widetilde{\boldsymbol{\mathcal{H}}}_1\boldsymbol{W}_{\widetilde{\boldsymbol{\mathcal{H}}}_1}^{(h)}, \boldsymbol{K}_m\boldsymbol{W}_{K_m}^{(h)}, \boldsymbol{V}_m\boldsymbol{W}_{V_m}) \cdot \boldsymbol{W}_{\widetilde{\mathcal{H}}_2}, \tag{20}$$

where $\boldsymbol{W}_{\widetilde{\boldsymbol{\mathcal{H}}}_1}^{(h)}$ and $\boldsymbol{W}_{K_m}^{(h)}$ are head-specific weights for keys and queries, $\boldsymbol{W}_{V_m}$ are value weights shared across all heads, $\boldsymbol{W}_{\widetilde{\mathcal{H}}_2}$ are used for final linear mapping of second phase. Furthermore, to facilitate residual concatenation, the features from each data source are initially aligned dimensionally using a linear transformation.

IV. Performance Evaluation

In this section, we describe the training procedure and eight benchmarks that contribute to the comparative analysis of our proposed framework. These models along with the proposed model are trained using 5-year data from 2016 to 2020, while data from 2021 are used for hyper-parameter tuning and data from 2022 is used for performance evaluation. The experimental results are then analyzed and discussed.

A. Training Procedure

To guarantee the fairness of the experimental evaluation process, all experiments are carried out using a single GPU. To mitigate overfitting during the training process, we implemented early stopping protocols with the patience parameter set to 5 epochs. All training procedures employ the Ranger [36] optimizer with a weight decay of 0.05 and deploy the cosine warmup strategy [37]. To enhance the performance of the model, we conducted an extensive search for optimal hyperparameters based on Optuna [38]. The chosen hyperparameters are shown in Table II.

TABLE II
Hyperparameters for SolarFusionNet

Items	Hyperparameters	Search Range
Batch Size	8	-
Input window	24	-
Learning rate	0.001	{0.0001, 0.1}
Kernel size for $conv_{1d}$	3	-
Kernel size for $conv_{3d}$	$5 \times 5 \times 5$	-
Hidden layer units	128	{64, 128, 256}
LSTM layer	3	{1, 3, 5}
ConvLSTM layer	3	{1, 3, 5}
GKConvLSTM layer	3	{1, 3, 5}
σ	1	-
Attention head	4	{1, 4, 8}
Output step	24	-
Dropout	0.5	{0.1, 0.9}

B. Benchmarks

To evaluate the effectiveness of the proposed model, we chose eight benchmarks for comparison, including several SOTA deep learning models designed specifically for prediction tasks: Cross Video Vision Transformer (CrossViViT) [11], Multiple Image Convolutional Long Short Term Memory Fusion Network (MICNN-L) [19], FEDformer [39], Autoformer [40], and TFT [30]. CrossViViT and MICNN-L are multi-modal fusion models for predicting solar irradiance that leverages the same inputs as SolarFusionNet. While the remaining three models can only be used to predict time series, they only use historical meteorological data as input. The hyperparameter configurations for each benchmark are established on the basis of guidelines from the literature. The SolarFusionNet$_{wof}$ and SolarFusionNet$_{ws}$ indicate SolarFusionNet without optical flow as inputs and without satellite images as inputs, respectively. Furthermore, we utilize the widely used smart persistence model as benchmark [1], which assumes CSI persist between time t and and time $t + \Delta t$,

$$\hat{y}_i^{t+\Delta t} = \text{GHI}_{\text{cs}}^{t+\Delta t} \cdot y_i^t. \tag{21}$$

Given the significance of time in temporal prediction scenarios, we also incorporate hour-of-day and day-of-year as auxiliary features for all of benchmarks.

C. Evaluation Metrics

To evaluate the performance of each predicion model, four metrics are used: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Normalized Root Mean Square Error (nRMSE), and Forecast Skill (S_f). S_f is compares the RMSE of the proposed model and the RMSE of the Smart Persistence model.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{22}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{23}$$

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{24}$$

$$\text{S}_\text{f} = 1 - \frac{\text{RMSE}_\text{Forecast}}{\text{RMSE}_\text{SmartPersistence}}, \tag{25}$$

where y_i is the observed values, $\hat{y}_i$ is the prediction values, $\bar{y}$ is the mean of the observed values. In the training process, RMSE is chosen as the loss function.

D. Comparison with benchmark models

To compare the models' performance, the prediction error metrics are calculated using only data from periods when $Z < 85°$ (i.e., data from night periods are excluded). The results of SolarFusionNet and the benchmarks are presented in Table III. As shown, SolarFusionNet significantly outperforms all benchmarks across the four subdatasets described in Section II, demonstrating the superiority of SolarFusionNet for ultra-short-term GHI forecasting. In terms of 10-minute-ahead prediction, SolarFusionNet performs the best at the CAB, PAL, and PAY sites, with RMSE of 70.45 W/m^2, 69.37 W/m^2, and 62.91 W/m^2, respectively, and $\mathrm{S_f}$ of 0.360, 0.370, and 0.315, respectively. The prediction accuracy of SolarFusionNet is slightly lower than that of CrossViViT only for the CNR station, with RMSE, nRMSE, MAE, and $\mathrm{S_f}$ of 62.98 W/m^2, 0.166, 37.02 W/m^2, and 0.301, respectively. The prediction accuracy of SolarFusionNet gradually improves as the prediction horizon increases at four BSRN stations, with $\mathrm{S_f}$ reaching a maximum of 0.540 for the 80-minute-ahead and 160-minute-ahead prediction horizons. Notably, SolarFusionNet significantly outperforms the sub-optimal model, CrossViViT, for the CAB, CNR, and PAL sites over a 240-minute forecast horizon. Specifically, the RMSE of the forecasts for these three sites is reduced by 7.72 W/m^2, 10.26 W/m^2, and 13.67 W/m^2, respectively. For PAY site, the prediction accuracy of SolarFusionNet is slightly less than that of CrossViViT. Specific data show that SolarFusionNet achieves a RMSE of 101.9 W/m^2, nRMSE of 0.280, MAE of 75.84 W/m^2, as well as a $\mathrm{S_f}$ of 0.374. Furthermore, it is evident that the prediction accuracy of SolarFusionNet at the four sites outperforms all the time-series prediction models.

In addition to statistical metrics, we employ a visualization method to further evaluate the predictive capabilities of the proposed model when compared to the benchmark models (CrossViViT, Autoformer, Smart Persistence). Figure 6 displays a comparison of sample time series for ground truth data alongside 240-minute-ahead forecasts at four stations. SolarFusionNet has been shown to demonstrate superior accuracy in predicting ramp events. At the CAB site on June 14, 2022, SolarFusionNet achieves a predicted RMSE of 117.8 W/m^2, while the other three benchmark models all have RMSEs greater than 150 W / m^2. Despite the occasional lag effect in forecasting, SolarFusionNet experiences such a phenomenon significantly less frequently than the benchmark models.

E. Analysis of multimodal input data

To more thoroughly explore the effects of satellite-derived optical flow signals and multi-channel satellite images on the prediction accuracy, we perform two comparison experiments at the four BSRN sites. The experimental results have been presented in Table III, the $\mathrm{S_f}$ of SolarFusionNet$_{wof}$ (without optical flow data) exhibits an average decline of 6.44%, 10.49%, 11.65%, and 12.75% across four prediction horizons, respectively. Similarly, the $\mathrm{S_f}$ of SolarFusionNet$_{ws}$ (without satellite data) shows more significant reductions, with averages of 11.84%, 12.34%, 13.61%, and 16.65% across the same prediction horizons, respectively, when compared to the SolarFusionNet model. The results further demonstrate the significant contribution of multi-channel satellite images and optical flow signals in improving the accuracy of solar irradiance prediction. Moreover, as the prediction horizon expands, the improvement in prediction accuracy becomes more pronounced.

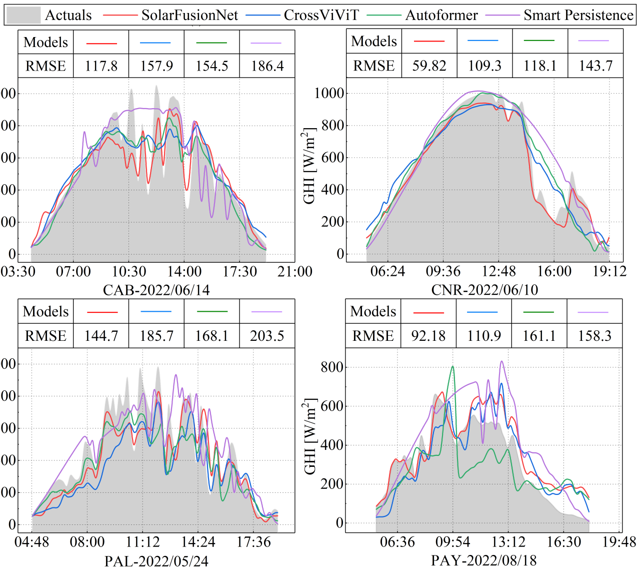


Fig. 6. Visualization results of SolarFusionNet and the benchmark models at four sites with 240-minute-ahead prediction details. The unit of RMSE is W/m^2.

F. Comparison under different weather conditions

Since cloud is the main atmospheric constitute that affects the available solar irradiance on the ground level, the predictive performance of the models are evaluated by comparing the forecasting accuracy under cloudy sky conditions [2], [14]. Specifically, the performance of models under cloudy skies reveals their ability to handle solar variability. In this research, we used the Bright-Sun clear-sky detection algorithm [41] to categorize the data from the four BSRN sites into clear-sky conditions and cloudy-sky conditions.

The performance of 240-minute-ahead GHI forecasting using SolarFusionNet, CrossViViT, Autoformer, and Smart Persistence are presented in Fig.7. In clear-sky conditions, the prediction RMSE of the four models are similar, and only Autoformer has a higher prediction RMSE at the PAY site. SolarFusionNet achieves the highest prediction accuracy at the CAB, CNR, and PAL sites under cloudy conditions, with a maximum RMSE reduction of 11.12 W/m^2 compared to the sub-optimal model, CrossViViT. For the PAY site, the prediction accuracy of SolarFusionNet is slightly lower than that of CrossViViT, with a difference in RMSE of only 0.73 W/m^2. The trend in prediction accuracy for the four models under cloudy conditions is consistent with the overall trend in prediction accuracy shown in Table III, which is due to the predominance of cloudy weather at the four sites throughout the year [20]. In general, the performance of SolarFusionNet is remarkable under cloudy conditions, especially at the CAB,

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT PREDICTION MODELS BASED ON FOUR BSRN STATIONS

Stations	Method	Prediction Horizons															
		10-minute-ahead				80-minute-ahead				160-minute-ahead				240-minute-ahead			
		RMSE	nRMSE	MAE	S_f	RMSE	nRMSE	MAE	S_f	RMSE	nRMSE	MAE	S_f	RMSE	nRMSE	MAE	S_f
CAB	Smart Persistence	110.1	0.321	62.75	-	170.8	0.532	111.1	-	181.1	0.572	120.5	-	170.7	0.545	108.4	-
	TFT[30]	76.29	0.245	46.14	0.307	112.2	0.360	74.01	0.343	130.7	0.419	89.42	0.279	140.9	0.453	99.78	0.175
	FEDformer[39]	70.58	0.226	40.73	0.359	94.83	0.304	56.49	0.445	104.1	0.334	64.15	0.426	112.9	0.363	71.1	0.339
	Autoformer[40]	73.96	0.237	42.81	0.328	87.81	0.282	**52.73**	0.486	97.14	0.311	60.17	0.464	102.5	0.331	65.03	0.399
	MICNN-L [19]	73.38	0.236	46.32	0.334	90.41	0.291	55.90	0.471	98.73	0.317	62.72	0.455	104.5	0.335	73.20	0.388
	CrossViViT[11]	72.29	0.232	42.41	0.343	86.16	0.277	55.82	0.495	96.64	0.310	67.15	0.467	102.4	0.330	72.46	0.400
	SolarFusionNet$_{wof}$	72.87	0.234	44.10	0.338	92.39	0.296	57.63	0.459	100.3	0.322	66.72	0.446	105.3	0.338	73.59	0.383
	SolarFusionNet$_{ws}$	74.19	0.238	47.39	0.326	94.18	0.302	61.32	0.449	106.5	0.342	72.89	0.412	107.8	0.346	84.29	0.368
	SolarFusionNet	**70.45**	**0.226**	**39.19**	**0.360**	**84.91**	**0.273**	53.25	**0.502**	**90.31**	**0.289**	**58.91**	**0.501**	**94.68**	**0.303**	**63.33**	**0.445**
CNR	Smart Persistence	90.09	0.235	48.16	-	163.3	0.428	100.9	-	181.5	0.476	116.6	-	173.1	0.454	107.9	-
	TFT[30]	73.63	0.193	45.54	0.183	109.6	0.288	71.18	0.329	123.3	0.324	82.61	0.321	130.7	0.344	89.02	0.245
	FEDformer[39]	71.79	0.188	49.78	0.203	92.23	0.242	62.42	0.435	100.4	0.264	67.69	0.447	102.9	0.271	70.72	0.405
	Autoformer[40]	64.24	0.169	35.96	0.286	88.93	0.234	53.33	0.455	97.92	0.257	59.81	0.461	105.5	0.277	64.93	0.390
	MICNN-L [19]	65.25	0.171	41.52	0.276	87.63	0.230	57.68	0.463	97.89	0.256	68.03	0.460	103.3	0.271	69.26	0.403
	CrossViViT[11]	**62.71**	**0.165**	**34.61**	**0.304**	86.92	0.228	**51.74**	0.468	95.37	0.251	**59.68**	0.475	100.9	0.265	**63.95**	0.417
	SolarFusionNet$_{wof}$	64.53	0.169	42.08	0.283	86.98	0.228	53.26	0.467	97.36	0.255	65.57	0.463	102.3	0.268	68.34	0.409
	SolarFusionNet$_{ws}$	66.71	0.175	40.25	0.260	88.11	0.231	60.59	0.460	99.62	0.261	71.38	0.451	105.9	0.277	74.36	0.388
	SolarFusionNet	62.98	0.166	37.02	0.301	**80.11**	**0.209**	53.57	**0.509**	**87.26**	**0.228**	63.58	**0.519**	**95.25**	**0.249**	67.37	**0.449**
PAL	Smart Persistence	110.2	0.326	59.70	-	166.3	0.493	104.6	-	179.8	0.532	113.6	-	171.4	0.508	104.5	-
	TFT[30]	83.98	0.249	51.43	0.238	114.4	0.339	76.34	0.312	135.9	0.403	94.221	0.244	149.9	0.445	107.98	0.125
	FEDformer[39]	74.49	0.221	46.73	0.324	87.51	0.259	61.73	0.473	103.5	0.307	76.62	0.424	116.8	0.347	88.28	0.318
	Autoformer[40]	75.42	0.223	43.53	0.316	90.23	0.268	58.65	0.457	99.58	0.295	67.03	0.446	106.3	0.315	73.55	0.380
	MICNN-L [19]	72.68	0.217	42.57	0.340	86.42	0.255	54.69	0.480	94.93	0.282	62.38	0.472	107.7	0.321	68.21	0.372
	CrossViViT[11]	70.23	0.208	37.84	0.365	83.00	0.246	51.66	0.500	93.62	0.278	61.17	0.479	104.3	0.309	69.08	0.391
	SolarFusionNet$_{wof}$	72.97	0.217	42.08	0.338	87.65	0.259	52.73	0.473	96.38	0.286	62.78	0.464	102.5	0.305	68.96	0.402
	SolarFusionNet$_{ws}$	74.21	0.221	44.39	0.327	89.36	0.265	56.79	0.463	100.2	0.297	66.38	0.443	105.5	0.314	75.36	0.384
	SolarFusionNet	**69.37**	**0.205**	**35.79**	**0.370**	**77.62**	**0.230**	**43.64**	**0.533**	**84.31**	**0.251**	**49.24**	**0.531**	**90.63**	**0.268**	**55.58**	**0.471**
PAY	Smart Persistence	91.86	0.252	45.52	-	159.7	0.438	91.97	-	172.7	0.475	103.9	-	162.8	0.447	96.30	-
	TFT[30]	74.95	0.206	43.24	0.184	112.5	0.309	70.81	0.296	131.8	0.363	88.58	0.234	147.6	0.406	105.1	0.093
	FEDformer[39]	67.80	0.186	38.27	0.261	85.17	0.234	57.33	0.467	100.3	0.276	71.83	0.416	111.9	0.308	83.10	0.312
	Autoformer[40]	74.62	0.205	43.39	0.188	117.2	0.322	75.83	0.266	136.4	0.374	92.66	0.208	148.4	0.408	104.9	0.088
	MICNN-L [19]	69.32	0.191	45.89	0.245	85.96	0.236	53.92	0.462	97.69	0.268	68.91	0.434	107.5	0.295	79.29	0.340
	CrossViViT[11]	68.01	0.187	38.93	0.259	81.67	0.225	54.03	0.489	92.16	0.254	63.31	0.464	**99.82**	**0.274**	**70.03**	**0.387**
	SolarFusionNet$_{wof}$	64.39	0.177	40.29	0.299	85.38	0.235	56.38	0.465	99.35	0.273	70.23	0.425	110.2	0.303	77.23	0.323
	SolarFusionNet$_{ws}$	66.59	0.183	42.36	0.275	87.26	0.240	58.63	0.453	103.2	0.283	72.22	0.402	112.8	0.309	81.22	0.307
	SolarFusionNet	**62.91**	**0.172**	**35.00**	**0.315**	**73.41**	**0.202**	**48.76**	**0.540**	**88.50**	**0.243**	**63.22**	**0.485**	101.9	0.280	75.84	0.374

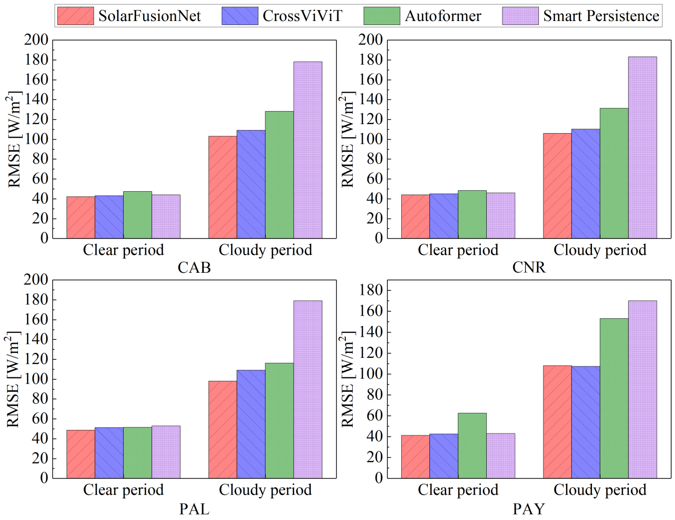


Fig. 7. The comparison of GHI prediction using SolarFusionNet, CrossViViT, Autoformer and Smart Persistence under clear and cloudy conditions at four BSRN stations.

CNR and PAL sites, which significantly improves the prediction accuracy.

To demonstrate the robustness of the SolarFusionNet model, the test datasets from the four BSRN sites are divided into four seasons according to meteorological criteria: spring (March to May), summer (June to August), fall (September to November), and winter (December to February). The 240-minute-ahead prediction RMSE of the four models for each season are presented in Fig. 8. SolarFusionNet has demonstrated superior performance and stability at CAB, CNR, and PAL sites, especially during summer months. At the PAY site, SolarFusionNet's prediction accuracy is slightly inferior to CrossViViT. The analysis reveals the overall superior GHI prediction performance and robustness of SolarFusionNet.

To comprehensively evaluate the solar irradiance prediction performance of SolarFusionNet under various weather conditions, we have performed a detailed interval division based on CSI. The comparison results presented in Fig. 9 clearly reflect the excellent performance of SolarFusionNet in different CSI intervals. In the interval of low CSI values (0-0.3), which represents cloudy or low-sunlight conditions, SolarFusionNet shows a significant advantage. In addition, SolarFusionNet performs well at all four BSRN sites. The results not only highlight SolarFusionNet's superior ability to handle extreme weather conditions, but demonstrate its strong potential to adapt to different geographic locations and diverse climatic conditions.

G. Uncertainty and robustness analysis

To comprehensively investigate the prediction capacity of SolarFusionNet, as suggested by Murphy and Winkler [42]

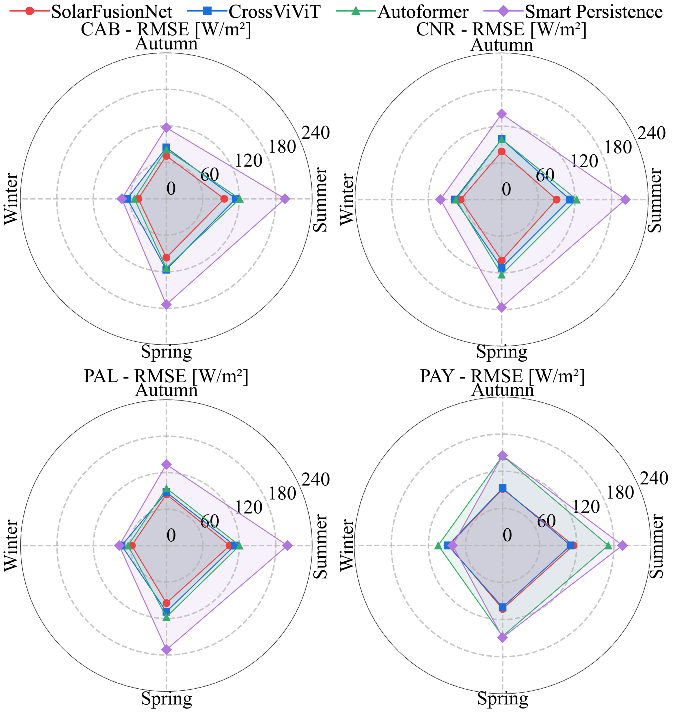


Fig. 8. The comparison of GHI prediction using SolarFusionNet, CrossViViT, Autoformer and Smart Persistence under different seasons at four BSRN stations.

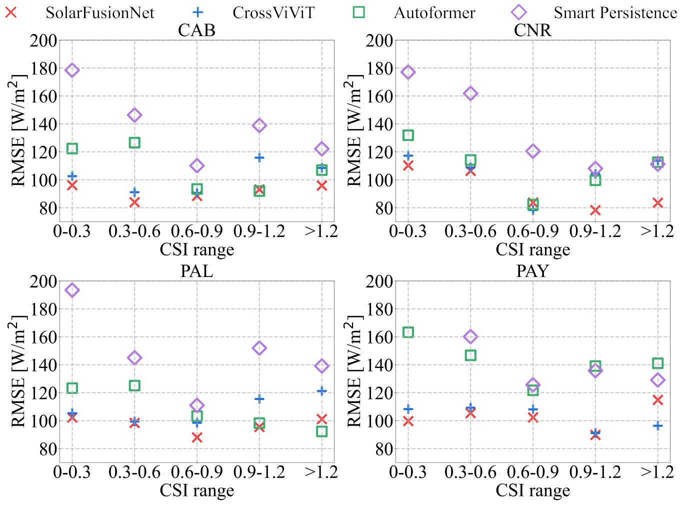


Fig. 9. The comparison of GHI prediction using SolarFusionNet, CrossViViT, Autoformer and Smart Persistence under different CSI intervals at four BSRN stations.

and Yang et al. [43], we utilize marginal distribution plots to analyze SolarFusionNet, CrossViViT, Autoformer, and Smart Persistence model in 240-minute-ahead GHI forecasts at CAB station. The joint and marginal distributions of measured and predicted GHI are depicted in Fig. 10. Compared to the benchmarks, the joint distribution of SolarFusionNet exhibits better alignment along the diagonal, which explains its smaller statistical errors. The probability density of SolarFusionNet is below the diagonal, indicating that the predicted GHI are generally lower than the measured values. A detailed analysis of the 2D kernel density contours for CAB reveals that the predicted values of SolarFusionNet are slightly below the identity line for high irradiance conditions, while the distribution of the predicted values tends to be near the identity line at lower irradiance levels.

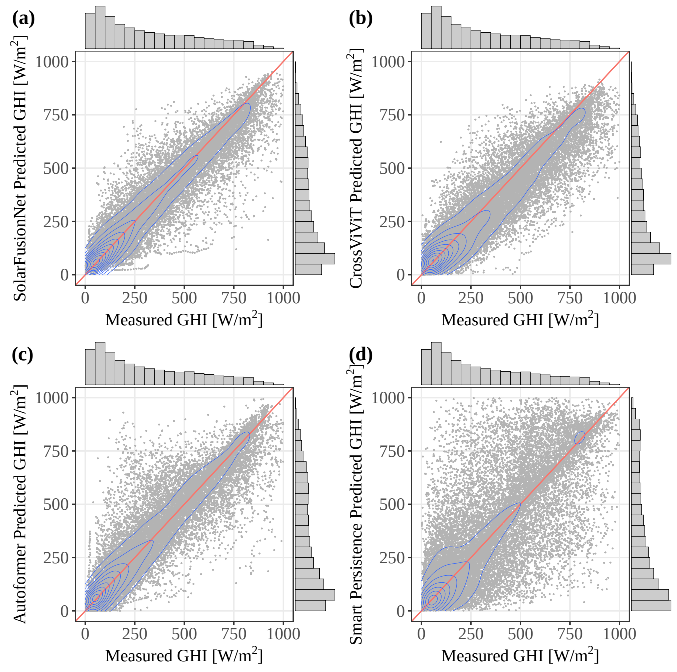


Fig. 10. Joint and marginal distributions of measured and predicted GHI using (a) SolarFusionNet, (b) CrossViViT, (c) Autoformer, and (d) Smart Persistence when evaluated at CAB station. The contour lines show the 2D kernel densities.

The histograms shown in Fig. 10 indicate the marginal distributions of observed values (on the top) and predicted values (on the right). The marginal distribution of the observed values shows that the peaks are located in the region of lower GHI, which suggests that the weather conditions at the CAB site in 2022 are skewed toward cloudy for most of the time. The predicted distributions for SolarFusionNet, CrossViViT, and Autoformer all exhibit the peak, but these peaks are shifted towards greater predictions compared to the measured values. In contrast, the predicted distribution of Smart Persistence undergoes a shift towards smaller predictions. Among the four models, SolarFusionNet exhibits the smallest shift, and the predicted value distribution closely matched the actual measurements.

Furthermore, Fig. 11 shows the conditional distributions to investigate the conditional dependence between observations and 240-minute-ahead predicted GHI using SolarFusionNet, CrossViViT, Autoformer and Smart Persistence. It is evident that when the measured GHI is smaller than 850 W/m^2, the peak of the local distribution of the predicted GHI of SolarFusionNet is more consistent with the diagonal, which means that the prediction accuracy of SolarFusionNet is higher. When the measured GHI exceeds 850 W/m^2, the deviation of the estimated value from the actual value increases, leading to a decrease in the accuracy of the prediction. In terms of the probability density of the prediction error, when $250\,\mathrm{W/m^2} < \mathrm{GHI} < 750\,\mathrm{W/m^2}$, the corresponding ridge diagrams are narrower, reflecting higher prediction accuracy.

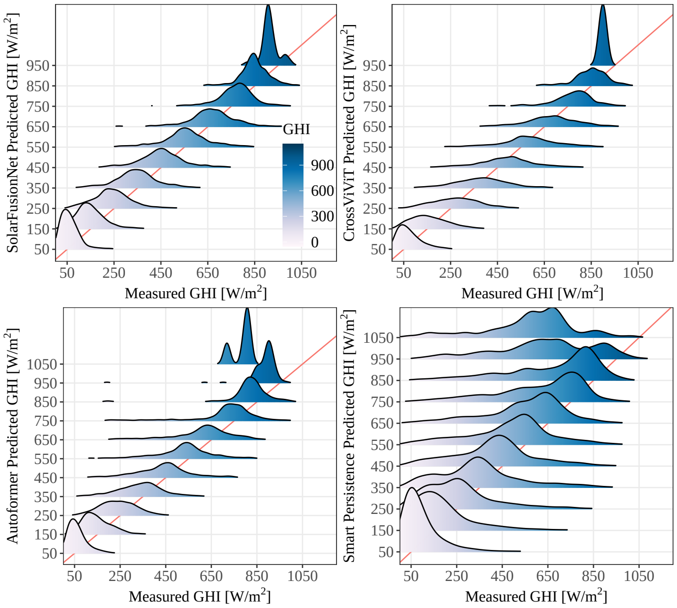


Fig. 11. Conditional distributions of 240-minute-ahead predicted GHI using SolarFusionNet, CrossViViT, Autoformer, and Smart Persistence.

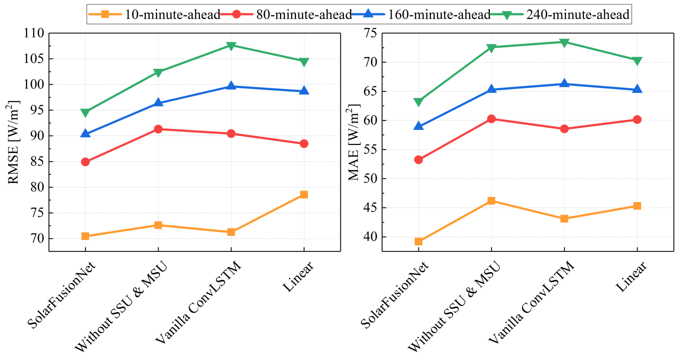


Fig. 12. Comparison of GHI prediction error metrics for three ablation experiments at CAB station.

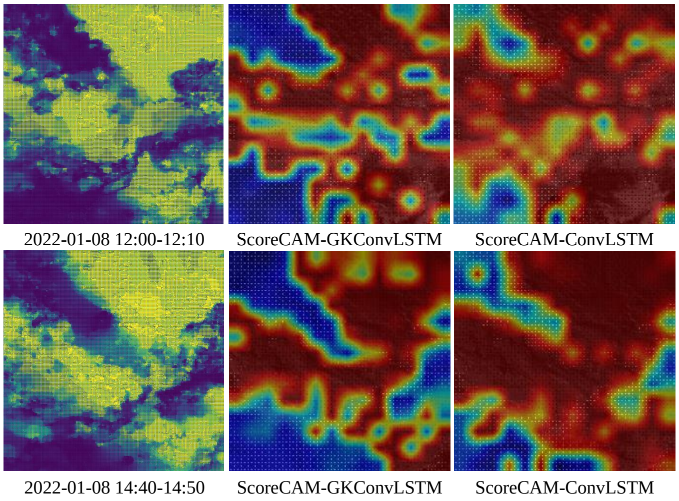


Fig. 13. Visualisation of the first layer GKConvLSTM and ConvLSTM on VIS 0.6 channel for optical flow features using ScoreCAM. The darker red colour of the heatmap represents higher significance of the feature.

H. Ablation analysis

To elucidate the significance of each component, we conducted three targeted ablation studies at the CAB station: we omitted the SSU and MSU modules, substituted GKConvLSTM with the vanilla ConvLSTM, and replaced the hierarchical multi-head cross-modal self-attention mechanism with a linear layer. Figure 12 shows the results of the ablation experiments in four different prediction horizons. It is shown that in the 10-minute-ahead and 80-minute-ahead forecasting, the SSU & MSU components and the multi-head cross-modal self-attention mechanism markedly affect the predictive accuracy of SolarFusionNet. Specifically, the MAE was reduced by 6.99 W/m^2 and 6.13 W/m^2, while the RMSE experienced reductions of 2.17 W/m^2 and 8.11 W/m^2, respectively. In contrast, the replacement of GKConvLSTM yielded more modest improvements, with a decrease of 3.39 W/m^2 in MAE and 0.81 W/m^2 in RMSE.

Within the scope of longer forecast horizons, the implementation of GKConvLSTM exhibits a pronounced enhancement in predictive accuracy. Compared to the performance of ConvLSTM in extracting features from optical flow signals, the error metrics of SolarFusionNet are significantly reduced by 7.35 W/m^2 and 10.16 W/m^2 for MAE and 9.31 W/m^2 and 12.99 W/m^2 for RMSE, respectively. This significant improvement can be primarily attributed to the adaptive weight mechanism of the Gaussian kernel, which exhibits superior proficiency in smoothing extraneous random noise that intensifies with the expansion of the predictive horizons. This enables the model to achieve enhanced stability of the longer-term predictive capacity. In this study, we also employ a convolutional feature insight technique, Score Class Activation Mapping (Score-CAM) [44], to visualize optical flow information captured by GKConvLSTM and ConvLSTM, respectively, as illustrated in Fig. 13. It is apparent that GKConvLSTM possesses a superior capability to pinpoint pivotal features within the optical flow signals with greater accuracy when compared with the vanilla ConvLSTM.

I. Variable significance analysis

We quantify the importance of each variable in different modalities by analyzing the weights ($\mathcal{V}_w$, $\mathcal{W}_w$) extracted from the automated multimodal feature selection units as described in Section III-A. The results of the analysis of the importance of the variables for the CAB station are depicted in Fig. 14. The variable importance analysis of the spectral channels indicates that the two visible channels, VIS 0.6 and VIS 0.8, hold the highest significance. This prominence is ascribed to the proficiency of the VIS 0.6 and VIS 0.8 channels in delineating the contours and configurations of cloud formations, coupled with their lower absorption of water vapor, rendering them particularly adept at detecting lower cloud strata and fog.

The variable importance analysis conducted on meteorological variables conclusively identifies historical CSI as the most influential contributor. Furthermore, BNI and DHI also demonstrate significant importance, with importance scores exceeding 0.1, which can be attributed to the direct physical relationship with GHI. Conversely, the relative importance of pressure (P), temperature (Temp), and relative humidity (RH)

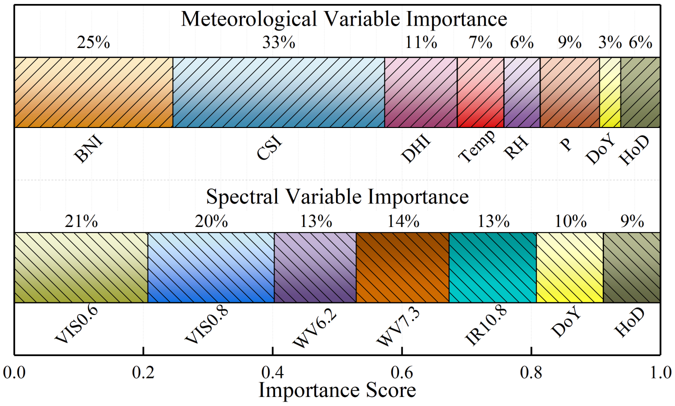


Fig. 14. Variable significance analysis at CAB station for 240-minute ahead forecasting.

are discerned to be comparatively marginal. Among these variables, pressure registers the highest importance score of 0.0921. The temporal features, day of the year (DoY) and hour of the day (HoD), exhibit minimal importance, a circumstance attributable to effective normalization of time-periodic factors in GHI prediction, achieved by the application of the clear-sky model.

V. Conclusion

The inherent uncertainty and intermittency of solar irradiance significantly impact the integration of PV power into the grid. Cloud movement is the primary cause of solar irradiance ramp events, therefore, integrating multispectral satellite images, derived optical flow information, and historical meteorological data is deemed an effective method to improve the accuracy of solar irradiance predictions. However, there have been limited studies on developing an end-to-end deep learning model that can simultaneously perform automatic selection and efficient fusion of multimodal features for regional solar irradiance prediction. This study introduces a deep learning model named SolarFusionNet, which uses a self-attention-based architecture that seamlessly integrates automatic multimodal feature selection and cross-modal fusion. Experimental results indicate that SolarFusionNet is capable of achieving SOTA prediction performance compared to advanced deep learning models. In the 240-minute-ahead prediction results, the S_f can reach 0.476, and RMSE and MAE are reduced by 13.67 W/m^2 and 13.50 W/m^2, respectively, compared to the suboptimal model.

References

[1] Y. Chu, M. Li, C. F. M. Coimbra, D. Feng, and H. Wang, "Intra-hour irradiance forecasting techniques for solar power integration: a review," *Iscience*, p. 103136, 2021.

[2] Y. Chu, Y. Wang, D. Yang, S. Chen, and M. Li, "A review of distributed solar forecasting with remote sensing and deep learning," *Renewable and Sustainable Energy Reviews*, vol. 198, p. 114391, 2024.

[3] D. Yang, W. Wang, and X. Xia, "A concise overview on solar resource assessment and forecasting," *Advances in Atmospheric Sciences*, vol. 39, no. 8, pp. 1239–1251, 2022.

[4] M. J. Mayer and G. Gróf, "Extensive comparison of physical models for photovoltaic power forecasting," *Applied Energy*, vol. 283, p. 116239, 2021.

[5] D. Yang, J. Kleissl, C. A. Gueymard, H. T. C. Pedro, and C. F. M. Coimbra, "History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining," *Solar Energy*, vol. 168, pp. 60–101, 2018.

[6] Y. Chu, M. Li, H. T. C. Pedro, and C. F. M. Coimbra, "A network of sky imagers for spatial solar irradiance assessment," *Renewable Energy*, vol. 187, pp. 1009–1019, 2022.

[7] S. Kharazi, N. Amjady, M. Nejati, and H. Zareipour, "A new closed-loop solar power forecasting method with sample selection," *IEEE Transactions on Sustainable Energy*, vol. 15, no. 1, pp. 687–698, 2024.

[8] S. Chen, C. Li, R. Stull, and M. Li, "Improved satellite-based intra-day solar forecasting with a chain of deep learning models," *Energy Conversion and Management*, vol. 313, p. 118598, 2024.

[9] H. S. Jang, K. Y. Bae, H.-S. Park, and D. K. Sung, "Solar power prediction based on satellite images and support vector machine," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.

[10] D. Yang, P. Jirutitijaroen, and W. M. Walsh, "Hourly solar irradiance time series forecasting using cloud cover index," *Solar Energy*, vol. 86, no. 12, pp. 3531–3543, 2012.

[11] O. Boussif, G. Boukachab, D. Assouline, S. Massaroli, T. Yuan, L. Benabbou, and Y. Bengio, "Improving day-ahead solar irradiance time series forecasting by leveraging spatio-temporal context," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[12] M. Nejati and N. Amjady, "A new solar power prediction method based on feature clustering and hybrid-classification-regression forecasting," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 2, pp. 1188–1198, 2021.

[13] H. Bouzgou and C. A. Gueymard, "Fast short-term global solar irradiance forecasting with wrapper mutual information," *Renewable energy*, vol. 133, pp. 1055–1065, 2019.

[14] S. Chen, C. Li, Y. Xie, and M. Li, "Global and direct solar irradiance estimation using deep learning and selected spectral satellite images," *Applied Energy*, vol. 352, p. 121979, 2023.

[15] S. Sharda, M. Singh, and K. Sharma, "Rsam: Robust self-attention based multi-horizon model for solar irradiance forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 2, pp. 1394–1405, 2020.

[16] D. P. Larson, M. Li, and C. F. M. Coimbra, "SCOPE: Spectral cloud optical property estimation using real-time GOES-R longwave imagery," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 2, p. 026501, 2020.

[17] J. Liu, H. Zang, L. Cheng, T. Ding, Z. Wei, and G. Sun, "A transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting," *Applied Energy*, vol. 342, p. 121160, 2023.

[18] X. Huang, J. Liu, S. Xu, C. Li, Q. Li, and Y. Tai, "A 3d convlstm-cnn network based on multi-channel color extraction for ultra-short-term solar irradiance forecasting," *Energy*, vol. 272, p. 127140, 2023.

[19] M. Ajith and M. Martínez-Ramón, "Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data," *Applied Energy*, vol. 294, p. 117014, 2021.

[20] A. Driemel, J. Augustine, K. Behrens, S. Colle, C. Cox, E. Cuevas-Agulló, F. M. Denn, T. Duprat, M. Fukuda, H. Grobe *et al.*, "Baseline surface radiation network (bsrn): structure and data description (1992–2017)," *Earth System Science Data*, vol. 10, no. 3, pp. 1491–1501, 2018.

[21] P. Ineichen, "A broadband simplified version of the solis clear sky model," *Solar Energy*, vol. 82, no. 8, pp. 758–762, 2008.

[22] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and future köppen-geiger climate classification maps at 1-km resolution," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[23] D. Yang, "Solardata package update v1. 1: R functions for easy access of baseline surface radiation network (bsrn)," *Solar Energy*, vol. 188, pp. 970–975, 2019.

[24] J. Schmetz, P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier, "An introduction to meteosat second generation (msg)," *Bulletin of the American Meteorological Society*, vol. 83, no. 7, pp. 977–992, 2002.

[25] W. Schumann, H. Stark, K. McMullan, D. Aminou, and H. Luhmann, "The msg system," *ESA bulletin*, pp. 11–14, 2002.

[26] D. M. A. Aminou, B. Jacquet, and F. Pasternak, "Characteristics of the meteosat second generation (msg) radiometer/imager: Seviri," in *Sensors, Systems, and Next-Generation Satellites*, vol. 3221. SPIE, 1997, pp. 19–31.

[27] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 Optical Flow Estimation," *Image Processing On Line*, vol. 3, pp. 137–150, 2013.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[30] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[31] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[33] S. Sinha, A. Garg, and H. Larochelle, "Curriculum by smoothing," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 653–21 664, 2020.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[36] L. Wright and N. Demeure, "Ranger21: a synergistic deep learning optimizer," *arXiv preprint arXiv:2106.13731*, 2021.

[37] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[38] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

[39] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International Conference on Machine Learning*, 2022, pp. 27 268–27 286.

[40] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 419–22 430, 2021.

[41] J. M. Bright, X. Sun, C. A. Gueymard, B. Acord, P. Wang, and N. A. Engerer, "Bright-sun: A globally applicable 1-min irradiance clear-sky detection model," *Renewable and Sustainable Energy Reviews*, vol. 121, p. 109706, 2020.

[42] A. H. Murphy and R. L. Winkler, "A general framework for forecast verification," *Monthly weather review*, vol. 115, no. 7, pp. 1330–1338, 1987.

[43] D. Yang, S. Alessandrini, J. Antonanzas, F. Antonanzas-Torres, V. Badescu, H. G. Beyer, R. Blaga, J. Boland, J. M. Bright, C. F. M. Coimbra *et al.*, "Verification of deterministic solar forecasts," *Solar Energy*, vol. 210, pp. 20–37, 2020.

[44] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

Tao Jing (Student Member, IEEE) received the B.S. degree in mechanical and automation engineering from Chang'an University, in 2018, and the M.S. degree in mechanical engineering from Northwestern Polytechnical University, in 2021.

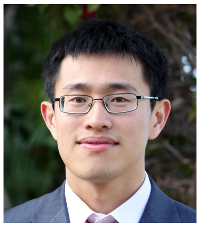
He is currently working toward the Ph.D. degree in the Department of Mechanical Engineering at The Hong Kong Polytechnic University. His research interests include solar irradiance forecasting, multimodal deep learning.

Shanlin Chen received the Ph.D. degree in Mechanical Engineering from The Hong Kong Polytechnic University, Hong Kong, China, in 2024.

He is currently a Postdoctoral Fellow with the Centre for Advances in Reliability and Safety, Hong Kong, China. His research interests include solar resource assessment and forecasting with machine learning, remote sensing, and numerical weather prediction.

David Navarro-Alarcon (Senior Member, IEEE) received the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong (CUHK), in 2014.

From 2014 to 2017, he worked as a Postdoctoral Fellow and then as a Research Assistant Professor at the T Stone Robotics Institute of CUHK. Since 2017, he has been with The Hong Kong Polytechnic University (PolyU), where he is currently an Associate Professor with the Department of Mechanical Engineering, and the Principal Investigator of the Robotics and Machine Intelligence Laboratory. His current research interests include perceptual robotics and control systems. Dr. Navarro-Alarcon currently serves as an Associate Editor of the IEEE TRANSACTIONS ON ROBOTICS.



Yinghao Chu received the Ph.D. degree in engineering sciences (mechanical engineering) from the University of California, San Diego, La Jolla, CA, USA, in 2015.

He is currently an Assistant Professor with the Department of Advanced Design and Systems Engineering, City University of Hong Kong, Hong Kong. He has been working in the domain of artificial intelligence (AI) for real-world application since 2011. His research interest includes hybrid AI, which simulates the attention and coordination mechanism of human intelligence to solve applicationorientated problems in real world, particularly in the areas of renewable forecast and smart manufacturing applications.

Mengying Li received her PhD in Mechanical and Aerospace Engineering from the University of California San Diego.

From 2018 to 2020, she served as a Postdoctoral Scholar at the Center for Energy Research at UC San Diego. Since 2020, she has been an Assistant Professor in the Department of Mechanical Engineering at The Hong Kong Polytechnic University. Currently, she is the Principal Investigator at the Renewable Energy Advancement Laboratory. Her research interests encompass solar resource assessment and forecasting, atmospheric radiative heat transfer, and the design of multigeneration systems.