SatSolarCast: A Flexible Framework for Multimodal Solar Irradiance Forecasting via Memory-Alignment Learning

Kuai Dai[1], Hui Su[1*], Chengxing Zhai[1], Huiwei Lin[2], Mingliang Bai[3]

[1]The Hong Kong University of Science and Technology
[2]The Chinese University of Hong Kong
[3]Harbin Institute of Technology
daikuai_hit@163.com, cehsu@ust.hk, cxzhai@ust.hk, felixlinhuiwei@gmail.com, mingliangbai@outlook.com

Abstract

Solar irradiance forecast aims to accurately estimate future solar irradiance based on historical data, playing a vital role in energy production and grid management. While ground-based station measurements provide local accuracy, geostationary satellites offer much broader environmental contexts, such as cloud coverage, which serves as a key factor for accurate forecasting. However, effectively integrating these multimodal observations remains a challenge, with existing methods suffering from inflexibility and high computational costs. To address this problem, we propose SatSolarCast, a flexible and efficient multimodal framework that introduces a memory alignment learning mechanism to integrate geostationary satellite data and historical irradiance observations. By preserving and recalling long-term spatiotemporal patterns from a specialized satellite memory bank, SatSolarCast enables effective guidance for both short- and long-term prediction. Additionally, SatSolarCast offers plug-and-play compatibility and can be incorporated into various forecasting architectures. Extensive experiments across four ground stations demonstrate that SatSolarCast substantially improves forecasting performance compared to prior methods with much lower computational costs.

Code — https://github.com/Applied-IAS/SatSolarCast

Introduction

Solar energy receives much attention due to its profound implications for energy security (Johnson et al. 2020; Capuano 2018), economic development (Junedi et al. 2022; Nguyen et al. 2018), environmental preservation (Hussain et al. 2023; Sharif et al. 2021), and climate change (Liu et al. 2023b; Khan et al. 2024). However, the amount of solar irradiance reaching the Earth's surface is not only modulated by diurnal and seasonal cycles but also affected by atmospheric conditions such as cloud cover. These fluctuations introduce substantial uncertainty into solar power production, making accurately forecasting solar irradiance vital for a wide range of applications, including energy generation and grid management.

In the past few years, data-driven approaches (Wang, Wang, and Su 2011; Chen et al. 2015; Alzahrani et al. 2017; Kumari and Toshniwal 2021; Zhou et al. 2021; Nie et al.

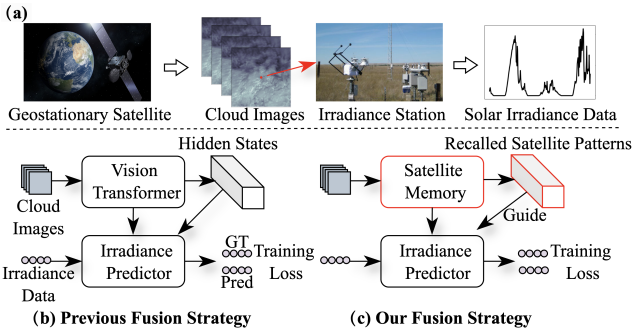

Figure 1: Illustration of our work. (a) The used multimodal data include geostationary satellite observations and station measurements. The red point in cloud images indicates the location of the ground station within the coverage area of the geostationary satellite data. (b) Previous fusion methods typically employ a Vision Transformer to model the spatiotemporal patterns of cloud image sequences and utilize its hidden states to guide solar irradiance forecast. However, these approaches are computationally expensive due to the high complexity brought by the self-attention mechanism. (c) In contrast, our fusion strategy introduces a satellite memory bank to preserve satellite spatiotemporal patterns and recall relevant patterns to guide solar irradiance prediction. This design significantly reduces computational costs and can be adapted into various solar irradiance prediction models.

Methods	R^2 ↑	RMSE ↓	Memory	Time	Params	FLOPs
CrossVIVIT	0.693	161.75	6.77GB	0.035s	144.94M	59.14G
LSTM-att+Sat	**0.727**	**152.00**	**1.06GB**	**0.0032s**	**10.09M**	**13.58G**

Table 1: Comparison of the Vision Transformer baseline, CrossVIVIT, and one case of our framework, LSTM-att+Sat, in average performance and costs for 3-hour solar irradiance forecasting, over four ground stations. The Memory and Time denote the training GPU memory demand and inference time for each sample, while the Params and FLOPs measure the parameter size and computational complexity, respectively.

*The corresponding author.

2023) have significantly advanced the field of solar irradiance forecast. As demonstrated in Figure 1 (a), the commonly used data include cloud images and solar irradiance data. Based on their usage patterns, the existing approaches are broadly categorized into three groups. First, utilizing historical solar irradiance data to predict future irradiance is a standard paradigm. These approaches (Hochreiter and Schmidhuber 1997; Luong, Pham, and Manning 2015; Jalali et al. 2021; Vaswani et al. 2017) model temporal patterns of irradiance time series using advanced architectures such as RNNs (Hochreiter and Schmidhuber 1997; Luong, Pham, and Manning 2015), CNNs (Jalali et al. 2021), and attention-based models (Vaswani et al. 2017; Zhou et al. 2021). Second, since cloud dynamics are a dominant factor affecting solar radiation variability, some studies (Le Guen and Thome 2020; Nie et al. 2024a; Xia et al. 2024) employ cloud images to achieve solar irradiance forecast. Specifically, these methods leverage historical cloud image sequences captured by sky cameras (Le Guen and Thome 2020; Nie et al. 2024a) or satellite scan imagers (Ahn et al. 2024; Sebastianelli et al. 2024; Xia et al. 2024) to conduct the ultra-short-term or long-term forecasting. Third, relying solely on irradiance data or cloud images limits prediction accuracy due to the lack of complementary context information. To address this, recent works (Liu et al. 2023a; Boussif et al. 2023) explore multimodal fusion strategies that integrate irradiance data with cloud imagery to improve prediction accuracy. However, as depicted in Figure 1 (b), these fusion methods typically employ a Vision Transformer (Dosovitskiy et al. 2021) to model the spatiotemporal patterns of cloud images and use it to guide solar irradiance forecasting, which leads to high computational complexity and limited scalability.

To address these limitations, we design an external memory-based framework that avoids modeling fine-grained dynamics with heavy spatiotemporal encoders. As shown in Figure 1 (c), we maintain a learnable satellite memory bank that preserves long-term satellite spatiotemporal patterns. During inference, short-term satellite observations are utilized as a query to recall possible satellite patterns for guiding the irradiance prediction, which enables both computational efficiency and flexibility. Table 1 quantitatively illustrates the advantages of our framework in both performance and efficiency. This design raises two key challenges: (1) how to employ the satellite memory bank to preserve spatiotemporal patterns of satellite cloud image sequences, especially long-term ones; (2) how to leverage learned patterns within the satellite memory bank to guide solar irradiance forecast. In addition, efficiently fusing the satellite patterns and irradiance features remains a non-trivial task.

To overcome the challenges, we propose SatSolarCast, an efficient and flexible multimodal forecasting framework. Specifically, to tackle the first and second challenges, SatSolarCast employs a two-phase training strategy, including a memory phase and a recall phase. In the memory phase, a spatiotemporal context encoder and a satellite memory bank are utilized to capture and preserve long-term spatiotemporal patterns from satellite sequences. In the recall phase, SatSolarcast recalls long-term patterns with short-term satellite sequences as queries to enhance solar irradiance prediction results. The two-phase mechanism enables SatSolarCast to effectively preserve long-term satellite patterns and recall long-term ones to guide forecasting. To further fuse the satellite patterns and irradiance features, we introduce a global spatial feature extractor and a spatiotemporal attention-based fusion method, which refine the fusion of satellite and irradiance data by providing global satellite context and modeling the dependencies between the two modalities, respectively. Moreover, SatSolarCast features plug-and-play compatibility, allowing it to be integrated with various solar irradiance prediction models. Comprehensive experiments across four ground stations demonstrate that SatSolarCast consistently enhances prediction accuracy across diverse forecasting models, achieving significant improvements in both 3- and 6-hour forecasting horizons with minor computational costs.

The main contributions of this work are summarized as follows:

- We proposed SatSolarCast, a flexible and efficient multimodal framework that introduces memory alignment learning to incorporate geostationary satellite data for improving solar irradiance forecast.
- The developed fusion elements and structures in SatSolarcast effectively integrate the spatiotemporal modalities of solar irradiance and satellite cloud data. Additionally, SatSolarCast offers plug-and-play compatibility and can be adapted to various solar irradiance prediction models.
- Extensive experiments over four ground stations confirm its effectiveness in consistently improving accuracy over different solar irradiance predictors, both for 3- and 6-hour prediction horizons with minor computational costs.

Related Work

Unimodal Methods

Solar Irradiance Forecast with Irradiance Data. Based on historical solar irradiance data, the methods focus on modeling temporal evolution patterns of solar irradiance sequences with machine learning techniques. Early work, such as (Wang, Wang, and Su 2011) introduces a backpropagation neural network to achieve solar irradiance forecast. To improve the forecast robustness, ensemble learning strategies such as Random Forest (Breiman 2001) and XGBoost (Chen et al. 2015) are later incorporated. With the development of deep learning techniques, advanced neural networks have gained much attention. For instance, (Alzahrani et al. 2017) applies LSTM to perform solar irradiance forecast and delivers promising results. (Sharda, Singh, and Sharma 2020) explores self-attention mechanisms for achieving long-term solar irradiance forecast. In addition, recent time series models (Xu et al. 2024, 2025) have brought insights for solar irradiance forecast. Informer (Zhou et al. 2021) proposes a sparse attention structure to capture long-dependency correlations in sequences. PatchTST (Nie et al. 2023) incorporates the patch technique from ViT (Dosovitskiy et al. 2021) and effectively captures local and global temporal patterns.

Solar Irradiance Forecast with Cloud Images. The pipelines for solar irradiance forecast with cloud images

generally follow two steps. First, extrapolating future cloud image frames based on historical observations, with spatiotemporal predictive models (Shi et al. 2015; Lee et al. 2019, 2021; Dai et al. 2023) such as ConvLSTM (Shi et al. 2015) and PredRNN (Wang et al. 2017), as well as more advanced frameworks like PhyDNet (Guen and Thome 2020) and MSTCGAN (Dai et al. 2022). Second, transforming the predicted cloud frames into cloud scores, which are then used to estimate solar irradiance, as demonstrated in studies such as (Le Guen and Thome 2020), (Paletta et al. 2022), (Mercier, Rahman, and Sabet 2023), (Xia et al. 2024), and (Nie et al. 2024b). From a data source perspective, the approaches leverage either sky images or satellite cloud images. The sky images are formed from ground-based cameras, offer high spatiotemporal resolution, and are suitable for ultra-short (seconds - 30 minutes) solar irradiance forecast. In contrast, satellite cloud images, produced by geostationary satellites using specialized wavebands, provide much broader spatial coverage and long-term monitoring, making them ideal for extended solar irradiance forecasts. Specifically, (Le Guen and Thome 2020) employs PhyDNet to predict solar irradiance based on sky images, while (Nie et al. 2024b) combines VideoGPT (Yan et al. 2021) and PhyDNet to achieve probabilistic solar irradiance forecast with sky images. (Xia et al. 2024) improves PredRNN to enhance cloud cover prediction at solar photovoltaic plants.

Despite the advancements, the methods either rely solely on historical solar irradiance time series or cloud image sequences, without incorporating the two types of data. In this work, we focus on enhancing solar irradiance forecast by integrating satellite cloud imagery, which is a key distinction from the aforementioned approaches.

Multimodal Methods

At present, the study about integrating solar irradiance data with cloud imagery for solar irradiance forecast is still in the preliminary stage. (Liu et al. 2023a) leverages sky images to improve ultra-short solar irradiance forecast. In particular, the approach adopts a dense optical flow and Vision Transformer (Dosovitskiy et al. 2021) to extract cloud motions of sky image sequences, to guide irradiance prediction. Then (Boussif et al. 2023) leverages satellite data to improve long-term solar irradiance forecast. The method similarly introduces a Vision Transformer to encode spatiotemporal patterns of satellite sequences and utilizes three other Transformer-based modules to extract temporal patterns and incorporate the two modalities. Though the approaches bring valuable insights for integrating solar irradiance measurements and cloud images, the high computation cost and low flexibility are the main concerns. To address these limitations, this work aims to develop an efficient and adaptable framework that leverages satellite data to enhance solar irradiance forecast.

Problem Definition

As for a standard solar irradiance forecast, given the k historical solar irradiance observations $X = \{X_1, X_2, \cdots X_k\} \in \mathbb{R}^{k\times 1}$, a prediction model $\mathcal{P}(\cdot)$ with parameters θ is expected to produce the t-length irradiance sequence $\hat{Y} = \{\hat{X}_{k+1}, \hat{X}_{k+2}, \cdots, \hat{X}_{k+t}\} \in \mathbb{R}^{t\times 1}$, which should be as similar to the ground-truth observations $Y = \{X_{k+1}, X_{k+2}, \cdots, X_{k+t}\} \in \mathbb{R}^{t\times 1}$ as possible. Formally, the procedure is defined as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{P}(X;\theta), Y). \tag{1}$$

The θ^* is the optimal parameters of model $\mathcal{P}(\cdot)$, and the $\mathcal{L}(\cdot)$ denotes the loss function to supervise the training procedure of the prediction model $\mathcal{P}(\cdot)$.

In this work, we aim to enhance the accuracy of solar irradiance forecast with geostationary satellite sequence $S = \{S_1, S_2, \cdots S_k\} \in \mathbb{R}^{k\times c\times H\times W}$. The c is the channel of satellite data at each time step, and H and W represent the height and width of the satellite image at each time step. Hence, this task can be redefined as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{P}(X; S;\theta), Y). \tag{2}$$

From the perspective of data dimensions, X can be considered as a sequence of points, while S represents an image sequence. Therefore, this task can be viewed as a multi-modal fusion and prediction problem.

Methodology

Training and Inference Flow of SatSolarCast

Training Flow As shown in Figure 2, the training procedure of SatSolarCast consists of two phases, the memory phase and the recall phase. Here, we briefly outline the workflow shared by the two phases. First, the short-term satellite sequence $S^s = \{S_1, S_2, \cdots S_k\} \in \mathbb{R}^{k\times c\times H\times W}$ is fed into a global spatial feature extractor $GST(\cdot)$ to obtain global spatial features $F_{\text{spatial}} \in \mathbb{R}^{k\times c}$. Second, F_{spatial} is concatenated with the historical solar irradiance sequence $X = \{X_1, X_2, \cdots X_k\} \in \mathbb{R}^{k\times 1}$, then passed into the solar irradiance predictor $\mathcal{P}(\cdot)$, to produce the hidden states $H \in \mathbb{R}^{k\times d}$, where d denotes the embedding size of hidden states at each time step. Third, the encoded hidden states $H \in \mathbb{R}^{k\times d}$ and long-term satellite patterns M^l (in the memory phase) or short-term ones M^s (in the recall phase) are fused through a spatiotemporal attention fusion mechanism $STatt(\cdot)$ to produce the predicted solar irradiance sequence $\hat{Y} = \{\hat{X}_{k+1}, X_{k+2}, \cdots, \hat{X}_{k+t}\} \in \mathbb{R}^{t\times 1}$. The procedure of the two phases is formally defined as follows:

Memory Phase: Preserve patterns of long-term satellite sequence M^l with a learnable satellite memory bank:

$$S^l = \{S_1, \cdots, S_{k+1}, \cdots, S_{k+t}\} \in \mathbb{R}^{(k+t)\times c\times H\times W}, \tag{3}$$

$$M^l = \mathcal{M}(STCE(S^l); \text{Update}(\epsilon)), \tag{4}$$

$$\hat{Y} = STatt(\mathcal{P}(GST(S^s) \oplus X;\theta), M^l). \tag{5}$$

Recall Phase: The short-term satellite sequence S^s serves as a query to recall possible patterns M^s from the frozen satellite memory bank:

$$M^s = \mathcal{M}(STCM(S^s); \text{Frozen}(\epsilon)), \tag{6}$$

$$\hat{Y} = STatt(\mathcal{P}(GST(S^s) \oplus X;\theta), M^s). \tag{7}$$

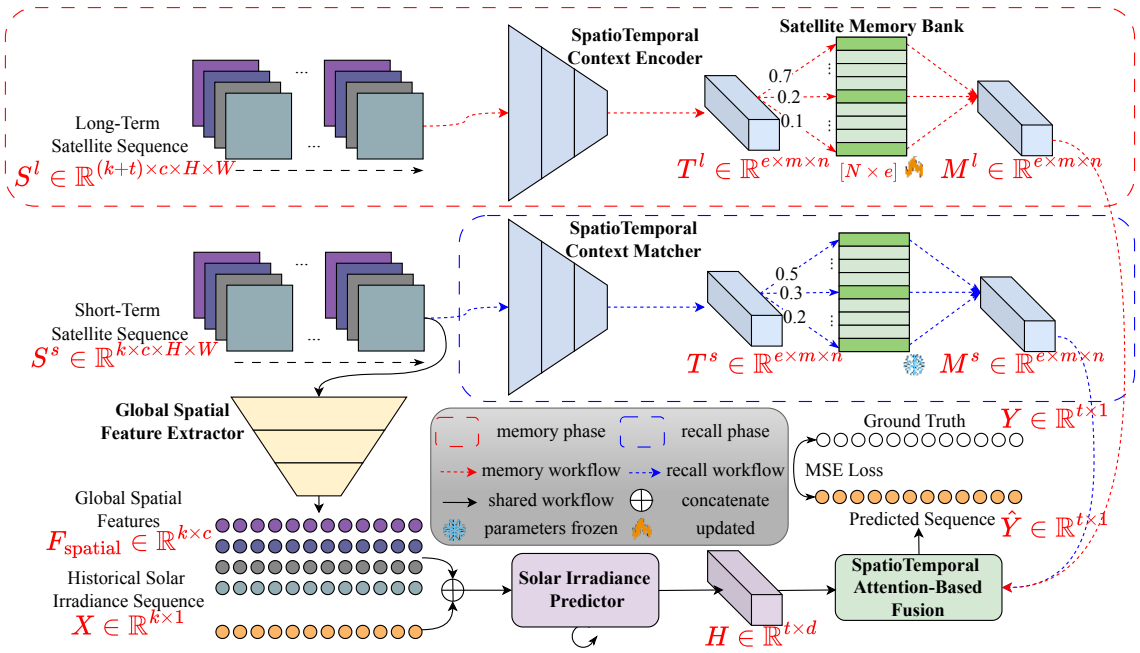

Figure 2: The overall layout of SatSolarCast. Its training process consists of two phases: the memory phase (red box) and the recall phase (blue box). In the memory phase, SatSolarCast focuses on preserving patterns within long-term satellite sequences, while in the recall phase, it recalls relevant satellite patterns to guide solar irradiance forecast with the short-term satellite sequences as queries. Notably, parameters of the satellite memory bank are updated in the memory phase while frozen during the recall phase. During the test procedure, SatSolarCast only conducts the second phase, to produce solar irradiance predictions.

The symbol $\oplus$ denotes concatenation, $\mathcal{M}(\cdot)$ represents the satellite memory bank, and ϵ denotes its parameters. $\mathrm{Update}(\cdot)$ and $\mathrm{Frozen}(\cdot)$ are parameter updated and frozen operations, respectively. $STCE(\cdot)$ and $STCM(\cdot)$ are the spatiotemporal context encoder and matcher, respectively.

Inference Flow During the test procedure, SatSolarCast merely conducts the recall phase to make predictions. Next, we introduce key modules of SatSolarCast in detail.

Model Design and Components

Global Spatial Feature Extractor This module first utilizes 2D convolution layers to aggregate spatial features of the short-term satellite sequence S^s. Then, it employs a 2D adaptive pooling layer and a fully-connected layer to capture global features F_{spatial}. The computation processes are as follows:

$$F_{\text{spatial}} = \mathrm{FC}[\mathrm{AdaPool}(\mathrm{Conv}(\mathrm{ReLU}[\mathrm{Conv}(S^s)]))]. \quad (8)$$

The $\mathrm{FC}[\cdot]$, $\mathrm{AdaPool}(\cdot)$, $\mathrm{Conv}(\cdot)$, and $\mathrm{ReLu}[\cdot]$ denote the fully-connected layer, adaptive pooling layer, convolution layer, and ReLU activation function, respectively.

Solar Irradiance Predictor The predictor accounts for modeling temporal evolution patterns from combined features $(F_{\text{spatial}} \oplus X)$ and transfers them into the hidden states $H \in \mathbb{R}^{t\times d}$ in an autoregressive manner, for producing the predicted solar irradiance sequence. The d is hidden states' size. In fact, the module can be replaced with any time series prediction model, such as GRU, LSTM, Transformer, or other more advanced models.

SpatioTemporal Context Encoder & Matcher The encoder is responsible for modeling the spatiotemporal patterns of satellite sequences. First, we use six 2D convolutional layers to downsample spatial features. Next, we design a simple yet effective temporal network, utilizing decomposed 3D convolutional layers, to capture temporal dynamics. Specifically, two 3D convolutional layers with a kernel size of (3, 1, 1) to extract temporal features, followed by another 3D convolutional layer with a kernel size of (1, 3, 3) to aggregate the features. Finally, we apply 3D adaptive pooling to reduce the temporal dimension to 1. In this manner, we can efficiently obtain the spatiotemporal patterns $T^l \in \mathbb{R}^{e\times m\times n}$ of long-term satellite sequences, which is calculated as follows:

$$T^l = \mathrm{3DAdaPool}(\mathrm{TemporalNet}(\mathrm{SpatialEnc}(S^l)). \quad (9)$$

$\mathrm{3DAdaPool}(\cdot)$ denotes the 3D adaptive pooling layer. $\mathrm{SpatialEnc}(\cdot)$ represents the encoder for downsampling the satellite sequences, consisting of the six cascaded convolutional layers. $\mathrm{TemporalNet}(\cdot)$ is the lightweight temporal network. In addition, the spatiotemporal context matcher accounts for extracting spatiotemporal patterns $T^s \in \mathbb{R}^{e\times m\times n}$ from short-term satellite sequence S^s, during the recall phase. It shares the same structure as the spatiotemporal context encoder.

Satellite Memory Bank The satellite memory bank is designed to memorize the extracted satellite spatiotemporal patterns. From the perspective of data formation, it is a matrix $\mathbf{M}$ with the shape of $N \times e$, which can be considered

as a set of N vectors with an embedding size of e. Each slot can memorize a kind of pattern. In this work, input satellite patterns T^l or T^s are first used to compute similarity scores with each slot in the satellite memory bank. Then, memorized satellite patterns M^l or M^s are obtained by computing a weighted sum of memory slots based on these similarity scores. Specifically, the computation process of the satellite memory bank is

$$\text{score}_\text{i} = \frac{\exp(cos(z, \mathbf{M}_i))}{\sum_{j=0}^{N-1}(\exp(cos(z, \mathbf{M}_j)))}, \tag{10}$$

$$out = \sum\nolimits_{i=1}^{N-1}(\text{score}_i \times \mathbf{M}_i). \tag{11}$$

$cos(\cdot)$ represents the cosine similarity score. $z \in \mathbb{R}^d$ denotes the partitioned tensor of input patterns, and out is corresponding output results from the satellite memory bank.

SpatioTemporal Attention-Based Fusion Through previous modules, we obtain hidden states $H \in \mathbb{R}^{t\times d}$ and satellite patterns $M^l \in \mathbb{R}^{e\times m\times n}$ or $M^s \in \mathbb{R}^{e\times m\times n}$ from the satellite memory bank. To effectively model the correlation between the solar irradiance data and satellite patterns, we design a spatiotemporal attention-based fusion approach. First, to match the channel size, we employ two deconvolution layers to project the satellite patterns M^l and M^s into $\hat{M}^l \in \mathbb{R}^{d\times h\times w}$ and $\hat{M}^s \in \mathbb{R}^{d\times h\times w}$, respectively. Second, the attention scores are computed using the hidden states H and projected satellite patterns $\hat{M}^l$ or $\hat{M}^s$. These scores are then applied to the hidden states to obtain enhanced features H_{enhanced}. Third, the enhanced features are concatenated with the hidden states and passed through a point-wise convolution layer to generate final fused features H_{final}. The computation process is

$$Q = H * W_Q; Q \in \mathbb{R}^{t\times d}, \tag{12}$$

$$K = (\hat{M}^l \text{ or } \hat{M}^s) * W_K; K \in \mathbb{R}^{d\times h\times w}, \tag{13}$$

$$K_{\text{flat}} = \text{reshape}(K); K_{\text{flat}} \in \mathbb{R}^{d\times(h\times w)}, \tag{14}$$

$$V = K_{\text{flat}}^T; V \in \mathbb{R}^{(h\times w)\times d}, \tag{15}$$

$$ATT = Q * K_{\text{flat}}; ATT \in \mathbb{R}^{t\times(h\times w)}, \tag{16}$$

$$ATT_i = \frac{\exp(ATT_i)}{\sum_{j=0}^{h\times w-1}\exp(ATT_j)}, \tag{17}$$

$$H_{\text{enhanced}} = ATT * V; H_{\text{enhanced}} \in \mathbb{R}^{t\times d}, \tag{18}$$

$$H_{\text{final}} = \text{Conv}_{1\times1}(H_{\text{enhanced}} \oplus H); H_{\text{final}} \in \mathbb{R}^{t\times d}. \tag{19}$$

The Q, K, and V denote the query, key, and value of the designed attention mechanism, respectively. $W_Q \in \mathbb{R}^{d\times d}$ and $W_K \in \mathbb{R}^{d\times d}$ are the projection matrices. ATT is the computed attention score. $\text{reshape}(\cdot)$ represents the dimension reshape operation, and K_{flat} is the corresponding flatten tensor. $\text{Conv}_{1\times1}(\cdot)$ is a convolution layer with a kernel size of 1×1. Finally, H_{final} is decoded into the predicted sequence $\hat{Y}$ with a fully-connected layer.

We utilize the mean squared error (MSE) loss to train SatSolarCast. Its training procedure is summarized in Algorithm 1. Notably, in the test procedure, SatSolarCast merely conducts the second phase, i.e., lines 9-11.

Algorithm 1: Training Procedure of SatSolarCast

Input: historical irradiance $X = \{X_1, ..., X_k\}$; ground-truth $Y = \{X_{k+1}, ..., X_{k+t}\}$; satellite observations $S = \{S_1, ..., S_{k+t}\}$

Output: predicted irradiance $\hat{Y} = \{\hat{X}_{k+1}, ..., \hat{X}_{k+t}\}$

1 initialize parameters Θ; learning rate lr; modules $GST(\cdot)$, $\mathcal{P}(\cdot)$, $STCE(\cdot)$, $STCM(\cdot)$, $\mathcal{M}(\cdot)$, $STatt(\cdot)$

2 **foreach** *batch* **do**

3 **Memory Phase:**

4 $S^s = \{S_1, ..., S_k\}$; $S^l = \{S_1, ..., S_{k+t}\}$;

5 $M^l = \mathcal{M}(STCE(S^l))$;

6 $\hat{Y} = STatt(\mathcal{P}(GST(S^s) \oplus X), M^l)$;

7 update parameters $\Theta \leftarrow \Theta - lr \cdot \nabla_\Theta \|Y - \hat{Y}\|_2^2$

8 **Recall Phase:**

9 freeze parameters ϵ of $\mathcal{M}(\cdot)$;

10 recalled patterns $M^s = \mathcal{M}(STCM(S^s))$;

11 $\hat{Y} = STatt(\mathcal{P}(GST(S^s) \oplus X), M^s)$;

12 update parameters $\Theta \leftarrow \Theta - lr \cdot \nabla_\Theta \|Y - \hat{Y}\|_2^2$

13 **end**

Experiments

Experimental Setup

Datasets. In this work, we utilize solar irradiance sequence data collected from multiple ground stations located in various countries such as the United States, Japan, and Australia to build and validate our approach. Complementary geostationary satellite observations from GOES-16 or Himawari-8 cover each station, including specialized spectral channels with wavelengths of 0.47 μm (visible), 0.86 μm (visible), 13.3 μm (infrared), as well as the solar zenith angle matrix. These data can be accessed through an open-source benchmark SolarCube (Li et al. 2024). To evaluate the forecasting ability under short- and long-term conditions, we extract 6-hour and 12-hour sequences, respectively. For the former, we use historical 3-hour observations (12 frames) to predict the next 3-hour sequence. For the latter, the goal is to predict the next 6-hour solar irradiance sequence (24 frames) using previous 6-hour observations. Our data division mainly follows the setting in (Li et al. 2024), one difference is that the station Cocos (COC) is removed due to insufficient observations to extract 12-hour sequences. In particular, we utilize the 14 stations, such as Chicago (born), Denver (fpk), Los_Angeles (dra), Chicago (sxf), $\cdots$, Tokyo (TAT), as training and validation data, while 4 stations, New York (psu), Denver (tbl), New York (LRC), and Tokyo (SAP) are used for testing.

Evaluation Metrics. We adopt two commonly used metrics in the solar irradiance forecasting domain, coefficient of determination (R^2) (Li et al. 2024) and root mean square error (RMSE) (Li et al. 2024), to evaluate the performance of the forecast models. The R^2 score quantifies the correlation between the predicted solar irradiance and ground truth. The RMSE measures the error between the prediction results and ground truth. Overall, a higher R^2 score and a lower value

Forecast Model	3-hour Forecasting								6-hour Forecasting							
	psu		tbl		LRC		SAP		psu		tbl		LRC		SAP	
	$R^2 \uparrow$	RMSE $\downarrow$	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Persistence	0.454	231.87	0.378	262.67	0.455	233.41	0.463	214.36	0.114	439.83	0.163	386.82	0.138	460.70	0.339	308.33
CrossVIVIT	0.690	157.50	0.658	179.39	0.722	153.06	0.666	156.26	0.607	187.97	0.537	220.76	0.661	178.29	0.561	196.22
LSTM	0.600	180.73	0.599	191.78	0.660	169.50	0.585	175.57	0.587	225.74	0.568	218.04	0.747	205.64	0.577	190.89
LSTM+Sat	0.729	147.17	0.674	172.88	0.756	144.26	0.715	143.39	0.661	192.79	0.611	201.07	0.796	190.04	0.645	183.11
Improvement (%)	**21.50**	**18.57**	**12.52**	**9.86**	**14.55**	**14.89**	**22.22**	**18.33**	**12.61**	**14.60**	**7.57**	**7.78**	**6.56**	**7.59**	**11.79**	**4.08**
LSTM-att	0.602	178.99	0.608	189.13	0.661	169.24	0.585	173.99	0.593	219.05	0.584	214.24	0.753	191.72	0.585	189.72
LSTM-att+Sat	0.735	145.64	0.680	170.16	0.752	147.59	0.712	144.00	0.665	182.81	0.611	205.20	0.797	180.50	0.646	179.04
Improvement (%)	**22.09**	**18.63**	**11.84**	**10.03**	**13.77**	**12.79**	**21.71**	**17.24**	**12.14**	**16.54**	**4.62**	**4.22**	**5.84**	**5.85**	**10.43**	**5.63**
Transformer	0.603	179.08	0.602	190.69	0.659	169.77	0.587	173.74	0.534	227.30	0.543	226.70	0.673	203.97	0.594	185.31
Transformer+Sat	0.716	153.01	0.671	176.30	0.728	151.46	0.684	157.34	0.630	197.23	0.583	208.78	0.711	193.16	0.657	181.07
Improvement (%)	**18.74**	**14.56**	**11.46**	**7.55**	**10.47**	**10.79**	**16.52**	**9.44**	**17.98**	**13.23**	**7.37**	**7.90**	**5.65**	**5.30**	**10.61**	**2.29**
PatchTST	0.548	191.57	0.533	206.64	0.604	182.59	0.539	183.46	0.496	244.15	0.509	229.33	0.681	222.16	0.504	210.46
PatchTST+Sat	0.714	151.97	0.657	176.73	0.733	150.41	0.690	150.17	0.607	200.81	0.572	216.05	0.693	185.32	0.611	195.28
Improvement (%)	**30.29**	**20.67**	**23.26**	**14.47**	**21.36**	**17.62**	**28.01**	**18.15**	**22.38**	**17.75**	**12.38**	**5.79**	**1.76**	**16.58**	**21.23**	**7.21**
Mamba	0.597	181.99	0.598	192.91	0.655	170.71	0.581	177.31	0.589	222.70	0.584	210.25	0.754	193.15	0.586	189.96
Mamba+Sat	0.736	145.42	0.677	172.83	0.751	145.07	0.695	149.33	0.647	183.77	0.601	205.08	0.784	173.29	0.623	185.61
Improvement (%)	**23.28**	**20.09**	**13.21**	**10.41**	**14.66**	**15.02**	**19.62**	**15.78**	**9.85**	**17.48**	**2.91**	**2.46**	**3.98**	**10.28**	**6.31**	**2.29**

Table 2: Quantitative results for 3- and 6-hour solar irradiance forecasting over four ground stations. LSTM+Sat is the abbreviation for the combination of LSTM and SatSolarCast. The bold highlights the improvements by SatSolarCast.

of RMSE indicate better performance.

Baseline Methods. We employ five typical single-modal forecasting approaches, Persistence (Trebing, Stanczyk, and Mehrkanoon 2021), LSTM (Hochreiter and Schmidhuber 1997), LSTM-att (Luong, Pham, and Manning 2015), Transformer (Vaswani et al. 2017), PatchTST (Nie et al. 2023), and Mamba (Gu and Dao 2024), and a recent multimodal method CrossVIVIT (Boussif et al. 2023) as baselines.

Implementation Details. For the global spatial feature extractor, the output channel sizes of the two convolution layers are 64 and 128, respectively. The embedding size c of the global features is 4. In the spatiotemporal context encoder and spatiotemporal context matcher, the spatial encoder contains six 2D convolution layers, with the output channel sizes of 64, 64, 128, 128, 256, and 512, respectively. In addition, the output channel sizes of the three 3D convolution layers are 256, 256, and 512, respectively. The satellite memory bank consists of 128 slots, each with a size of 512. In the spatiotemporal attention-based fusion method, the output channel sizes of the two deconvolution layers are 256 and 128, respectively. Although the solar irradiance predictor can be any arbitrary time series prediction model, its depth is set to 4, and the hidden state size is set to 128 for a fair comparison. We train SatSolarCast using the Adam optimizer with a learning rate of 2×10^{-4}. The experiments are conducted on a server with a CPU of Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80GHz and a RTX A800 GPU.

Overall Comparison

By analyzing Table 2, we have the following findings. First, SatSolarCast substantially improves forecasting accuracy overall for all unimodal baselines across four stations. Specifically, in the 3-hour overall evaluation of the four stations, the most significant improvement is achieved on the baseline PatchTST model, with a 24.04% increase in the R^2 score and a 17.57% reduction in RMSE. In the 6-hour overall evaluation, the highest improvement in the R^2 score, 21.92%, and in the RMSE score, 12.07%, is again achieved on PatchTST. Second, compared with the multimodal baseline CrossVIVIT, the combinations of SatSolarCast and the four unimodal baselines, especially the LSTM-att+Sat, significantly outperform CrossVIVIT at almost all validation settings. Moreover, at Table 1, we have demonstrated the notable superiority of our method in terms of computational costs. The observations consistently suggest flexibility, effectiveness, and efficiency of SatSolarCast.

To investigate forecast ability of SatSolarCast w.r.t the lead time, we present the MSE curves of the best case LSTM-att+Sat in Figure 3. First, SatSolarCast significantly improves the 3-hour forecast accuracy across all the baselines, and the advantages become more obvious as the lead time goes by. Second, SatSolarCast also substantially improves the 6-hour forecast. An interesting observation is that the MSE scores decrease during certain periods, likely due to the absence of sunlight. Additionally, around the 6th hour, a slight performance degradation is observed. This is reasonable, as without sunlight, the satellite patterns offer limited guidance for solar irradiance prediction.

To intuitively compare improvements provided by SatSolarCast, Figure 4 shows the visualization results of 3- and 6-hour solar irradiance forecasts, respectively. Notably, the X-axis is the ground truth while the Y-axis represents pre-

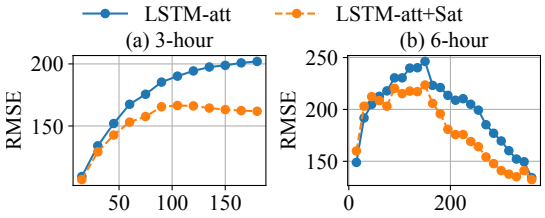

Figure 3: Comparison of MSE curves across four ground stations for both 3- and 6-hour solar irradiance forecasts.

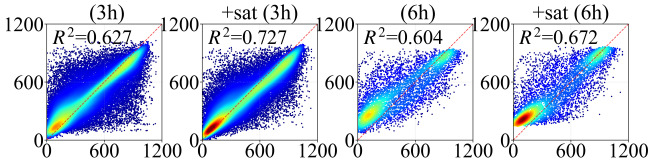

Figure 4: The comparison of corresponding visualization.

dicted values. The deeper color denotes the higher density of sample points for solar irradiance. We see that the SatSolarCast effectively improves the prediction accuracy, especially for the low irradiance (around 200 W/m^2) forecast. All the observations again demonstrate the superiority of SatSolarCast.

Ablation Studies

To verify the effectiveness of key components and structures in the proposed SatSolarCast, we conduct ablation studies by answering the following four questions.

(1) Does the memory alignment learning algorithm work for SatSolarCast? The core idea of SatSolarCast is to align short-term satellite spatiotemporal patterns with long-term ones. To achieve this goal, we adopt the two-phase procedure to train SatSolarCast. To verify the effectiveness of the training strategy, we remove the recall phase in the training procedure, resulting in a variant without the recall mechanism (w/o recall). As shown in Table 3, excluding the recall phase leads to a significant decline in performance across various solar irradiance prediction models. Specifically, without the recall phase, LSTM experiences a 9.68% decrease in the R^2 score and a 17.18% increase in the RMSE score. The other three models show similar performance degradation. The observations indicate the effectiveness of the memory alignment learning mechanism.

(2) Are the global spatial features & spatiotemporal attention-based fusion approach important for solar irradiance forecast? First, to validate the necessity of global spatial features, we remove the global spatial feature extractor (GSFE) and obtain a new variant (w/o GSFE). As a result, we observe that the absence of global spatial features leads to a performance decline. Second, to verify the effectiveness of the fusion method, we replace the spatiotemporal attention-based fusion approach with a simple fusion method and deliver a new variant (w/o STABF). In the variant, the spatiotemporal patterns recalled from the satellite memory bank are first fed into a 2D adaptive pooling layer and then are added to hidden states produced by the solar

Variants	LSTM +Sat		LSTM-att +Sat		Transformer +Sat		PatchTST +Sat	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
w/o recall	0.656	178.21	0.660	172.56	0.648	178.43	0.690	167.29
w/o GSFE	0.702	160.56	0.705	158.28	0.681	174.12	0.645	176.49
w/o STABF	0.723	153.97	0.726	153.12	0.705	**159.74**	0.703	159.66
w/ all	**0.726**	**152.18**	**0.727**	**152.00**	**0.707**	159.79	**0.707**	**157.52**

Table 3: The ablation results for 3-hour forecasting.

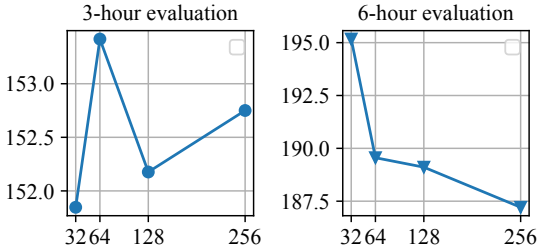

Figure 5: The RMSE curves w.r.t the slot numbers.

Costs (6-hour)	Parameters Increment	FLOPs Increment
	8.96M	26.98G

Table 4: The additional costs produced by SatSolarCast.

irradiance predictor. In fact, the simple fusion method is a special case of our spatiotemporal attention-based fusion approach, where attention scores are equal across the space-time dimensions. Similarly, without the fusion method, consistent performance decline is again observed. The comparison results highlight usefulness of the two structures.

(3) How does the satellite memory bank size affect the forecasting performance? Here, we employ the simplest combination, namely, LSTM and SatSolarCast, to conduct experiments. As shown in Figure 5, we observe that an appropriate slot number is important for forecasting accuracy. Notably, the slot number of 128 achieves a good balance for both 3- and 6-hour forecasting.

(4) What are the additional costs produced by SatSolarCast? As shown in Table 4, SatSolarCast merely takes an additional 8.96M parameters and 26.98G FLOPs cost for 6-hour forecasting. Considering performance benefits brought by SatSolarCast, this quantitative analysis highlights SatSolarCast's efficiency in improving solar irradiance forecast.

Conclusion

In this work, we propose a flexible and efficient multimodal solar irradiance forecasting framework SatSolarCast, which can significantly improve prediction accuracy across various model architectures, through memory-alignment learning to incorporate satellite imagery and irradiance data. Comprehensive experiments and analysis demonstrate the superiority of SatSolarcast and the effectiveness of its elements.

Acknowledgments

This work was supported in part by the Hong Kong Jockey Club Charities Trust (FA123 and P0413), the Innovation and Technology Commission project ITP/047/23LP (P0456) managed by the Hong Kong Logistics and Supply Chain MultiTech R&D Centre, and the State Key Laboratory scheme under Innovation and Technology Commission (ITC-SKLCRCC26EG01).

References

Ahn, H.; Yu, J.; Ko, J.; and Yeom, J.-M. 2024. Enhanced short-term prediction of solar radiation using HRNet model with geostationary satellite data. *IEEE Geoscience and Remote Sensing Letters*.

Alzahrani, A.; Shamsi, P.; Dagli, C.; and Ferdowsi, M. 2017. Solar irradiance forecasting using deep neural networks. *Procedia Computer Science*, 114: 304–313.

Boussif, O.; Boukachab, G.; Assouline, D.; Massaroli, S.; Yuan, T.; Benabbou, L.; and Bengio, Y. 2023. Improving* day-ahead* solar irradiance time series forecasting by leveraging spatio-temporal context. *Proceedings of the Advances in Neural Information Processing Systems*, 36: 2342–2367.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Capuano, L. 2018. International energy outlook 2018 (IEO2018). *US Energy Information Administration (EIA): Washington, DC, USA*, 2018: 21.

Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4): 1–4.

Dai, K.; Li, X.; Ma, C.; Lu, S.; Ye, Y.; Xian, D.; Tian, L.; and Qin, D. 2023. Learning spatial–temporal consistency for satellite image sequence prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

Dai, K.; Li, X.; Ye, Y.; Feng, S.; Qin, D.; and Ye, R. 2022. MSTCGAN: Multiscale time conditional generative adversarial network for long-term satellite image sequence prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First conference on language modeling*.

Guen, V. L.; and Thome, N. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11474–11484.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hussain, B.; Naqvi, S. A. A.; Anwar, S.; and Usman, M. 2023. Effect of wind and solar energy production, and economic development on the environmental quality: Is this the solution to climate change? *Gondwana Research*, 119: 27–44.

Jalali, S. M. J.; Ahmadian, S.; Kavousi-Fard, A.; Khosravi, A.; and Nahavandi, S. 2021. Automated deep CNN-LSTM architecture design for solar irradiance forecasting. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1): 54–65.

Johnson, O. W.; Han, J. Y.-C.; Knight, A.-L.; Mortensen, S.; Aung, M. T.; Boyland, M.; and Resurrección, B. P. 2020. Intersectionality and energy transitions: A review of gender, social equity and low-carbon energy. *Energy Research & Social Science*, 70: 101774.

Junedi, M.; Ludin, N.; Hamid, N.; Kathleen, P.; Hasila, J.; and Affandi, N. A. 2022. Environmental and economic performance assessment of integrated conventional solar photovoltaic and agrophotovoltaic systems. *Renewable and Sustainable Energy Reviews*, 168: 112799.

Khan, A.; Anand, P.; Garshasbi, S.; Khatun, R.; Khorat, S.; Hamdi, R.; Niyogi, D.; and Santamouris, M. 2024. Rooftop photovoltaic solar panels warm up and cool down cities. *Nature Cities*, 1(11): 780–790.

Kumari, P.; and Toshniwal, D. 2021. Deep learning models for solar irradiance forecasting: A comprehensive review. *Journal of Cleaner Production*, 318: 128566.

Le Guen, V.; and Thome, N. 2020. A deep physical model for solar irradiance forecasting with fisheye images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 630–631.

Lee, J.-H.; Lee, S. S.; Kim, H. G.; Song, S.-K.; Kim, S.; and Ro, Y. M. 2019. Mcsip net: Multichannel satellite image prediction via deep neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3): 2212–2224.

Lee, S.; Kim, H. G.; Choi, D. H.; Kim, H.-I.; and Ro, Y. M. 2021. Video Prediction Recalling Long-term Motion Context via Memory Alignment Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3054–3063.

Li, R.; Xie, Y.; Jia, X.; Wang, D.; Li, Y.; Zhang, Y.; Wang, Z.; and Li, Z. 2024. SolarCube: An Integrative Benchmark Dataset Harnessing Satellite and In-situ Observations for Large-scale Solar Energy Forecasting. In *Proceedings of the Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

Liu, J.; Zang, H.; Cheng, L.; Ding, T.; Wei, Z.; and Sun, G. 2023a. A Transformer-based multimodal-learning framework using sky images for ultra-short-term solar irradiance forecasting. *Applied Energy*, 342: 121160.

Liu, L.; He, G.; Wu, M.; Liu, G.; Zhang, H.; Chen, Y.; Shen, J.; and Li, S. 2023b. Climate change impacts on planned supply–demand match in global wind and solar energy systems. *Nature Energy*, 8(8): 870–880.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1412–1421.

Mercier, T. M.; Rahman, T.; and Sabet, A. 2023. Solar irradiance anticipative transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2065–2074.

Nguyen, S.; Peng, W.; Sokolowski, P.; Alahakoon, D.; and Yu, X. 2018. Optimizing rooftop photovoltaic distributed generation with battery storage for peer-to-peer energy trading. *Applied Energy*, 228: 2567–2580.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations*.

Nie, Y.; Zelikman, E.; Scott, A.; Paletta, Q.; and Brandt, A. 2024a. Skygpt: Probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained videogpt. *Advances in Applied Energy*, 14: 100172.

Nie, Y.; Zelikman, E.; Scott, A.; Paletta, Q.; and Brandt, A. 2024b. SkyGPT: Probabilistic Ultra-short-term Solar Forecasting Using Synthetic Sky Images from Physics-constrained VideoGPT. *Advances in Applied Energy*, 100172.

Paletta, Q.; Hu, A.; Arbod, G.; Blanc, P.; and Lasenby, J. 2022. SPIN: Simplifying Polar Invariance for Neural networks Application to vision-based irradiance forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 5182–5191.

Sebastianelli, A.; Serva, F.; Ceschini, A.; Paletta, Q.; Panella, M.; and Le Saux, B. 2024. Machine learning forecast of surface solar irradiance from meteo satellite data. *Remote Sensing of Environment*, 315: 114431.

Sharda, S.; Singh, M.; and Sharma, K. 2020. RSAM: Robust self-attention based multi-horizon model for solar irradiance forecasting. *IEEE Transactions on Sustainable Energy*, 12(2): 1394–1405.

Sharif, A.; Meo, M. S.; Chowdhury, M. A. F.; and Sohag, K. 2021. Role of solar energy in reducing ecological footprints: An empirical analysis. *Journal of Cleaner Production*, 292: 126028.

Shi, X. J.; Chen, Z. R.; Wang, H.; Yeung, D. Y.; Wong, W. K.; and Woo, W. C. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, 802–810.

Trebing, K.; Stanczyk, T.; and Mehrkanoon, S. 2021. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognition Letters*, 145: 178–186.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*, 5998–6008.

Wang, Y. B.; Long, M. S.; Wang, J. M.; Gao, Z. F.; and Yu, P. S. 2017. PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, 879–888.

Wang, Z.; Wang, F.; and Su, S. 2011. Solar irradiance short-term prediction model based on BP neural network. *Energy Procedia*, 12: 488–494.

Xia, P.; Zhang, L.; Min, M.; Li, J.; Wang, Y.; Yu, Y.; and Jia, S. 2024. Accurate nowcasting of cloud cover at solar photovoltaic plants using geostationary satellite images. *Nature Communications*, 15(1): 510.

Xu, X.; Chen, C.; Liang, Y.; Huang, B.; Bai, G.; Zhao, L.; and Shu, K. 2024. Sst: Multi-scale hybrid mamba-transformer experts for long-short range time series forecasting. *arXiv preprint arXiv:2404.14757*.

Xu, X.; Zhao, Y.; Philip, S. Y.; and Shu, K. 2025. Beyond numbers: A survey of time series analysis in the era of multimodal llms. *Authorea Preprints*.

Yan, W.; Zhang, Y.; Abbeel, P.; and Srinivas, A. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11106–11115.